

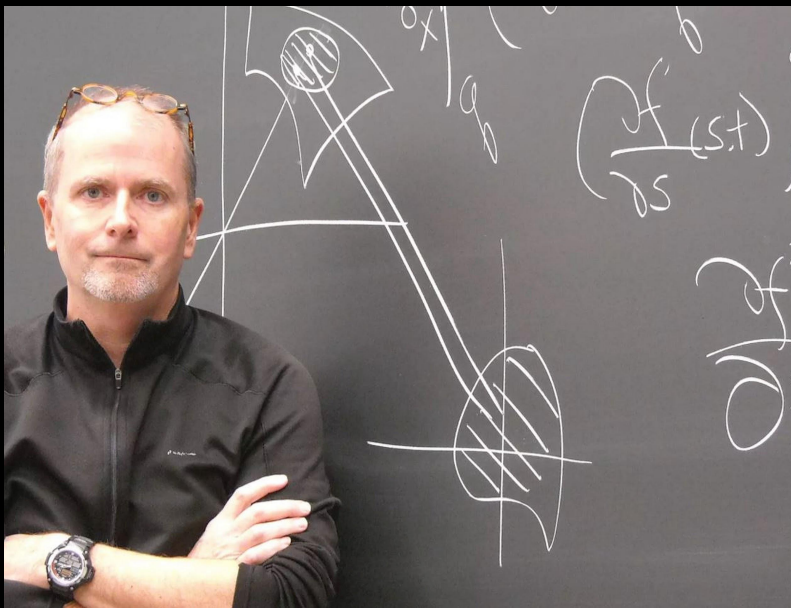
Machine Learning + Libraries “Look Book”



Machine Learning + Libraries Summit
Participant Project Lookbook
September 2019

extracting space/

the theory and application of convolutional neural nets &
deep learning in geospatial archives



john hessler

specialist in computational geography & geographic
information science

geography and map division

library of congress

jhes@loc.gov

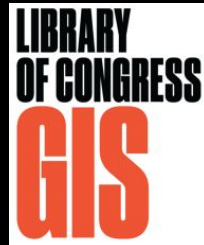
director, topology lab for the study of deep learning

lecturer in quantum theory, algorithms & computing

johns hopkins university

jhessle1@jhu.edu

<https://jhessler.net>

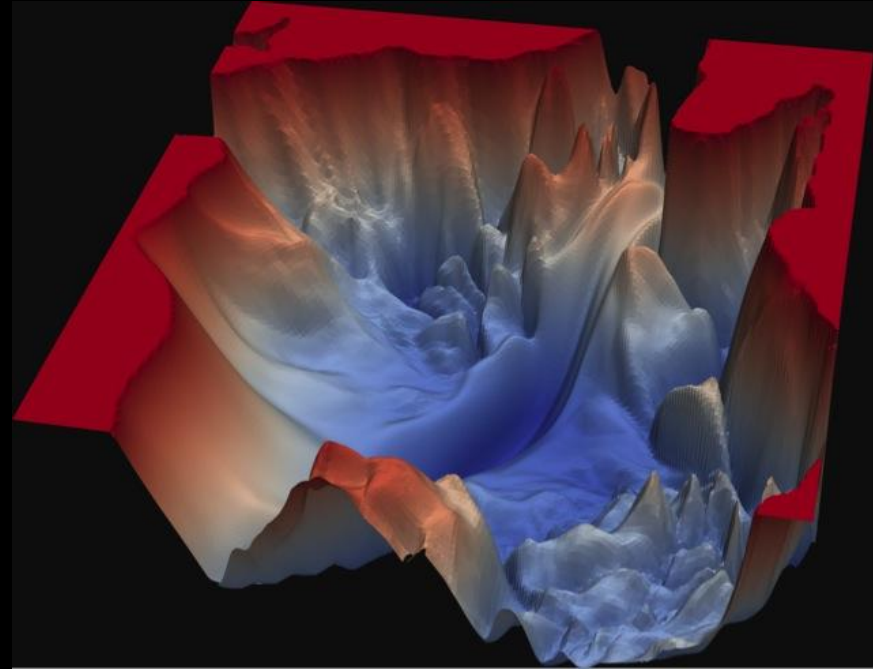


deep learning/feature extraction

- extracting spatial features from historic maps for use in GIS & geo-ai applications
 - convolutional neural networks
- stochastic gradient descent
 - backpropagation

OPEN QUESTIONS

- characterization & visualization of error landscapes
- how does all this actually work

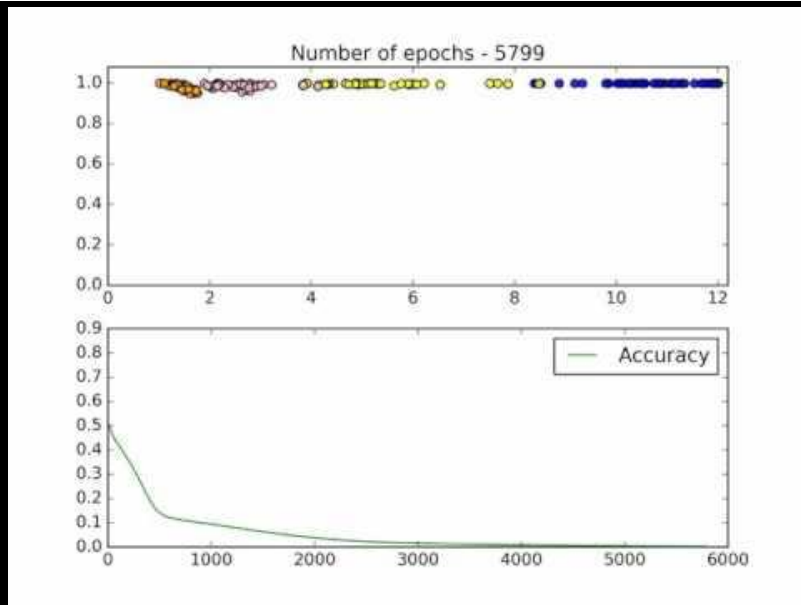
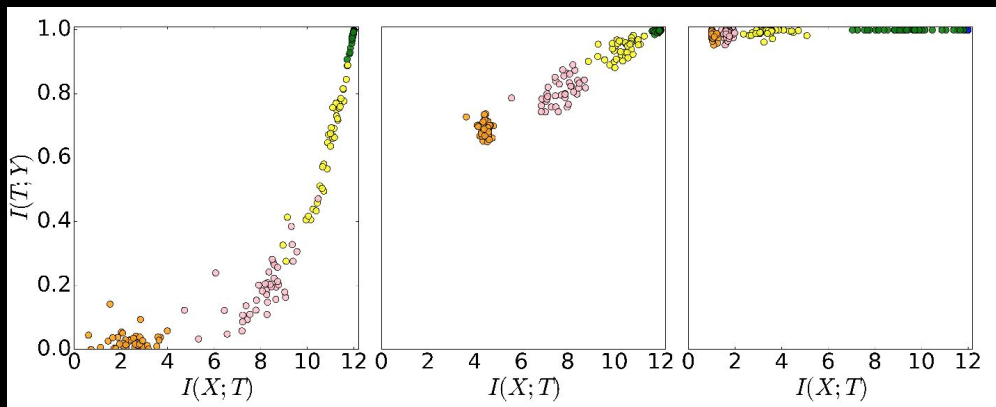


theories of deep learning/

- mutual Information

$$I(X;Y) \geq I(T_1;Y) \geq I(T_2;Y) \geq \dots \geq I(T_k;Y) \geq I(\hat{Y};Y)$$
$$H(X) \geq I(X;T_1) \geq I(X;T_2) \geq \dots \geq I(X;T_k) \geq I(X;\hat{Y})$$

- hidden layers modeled as markov chains
- phase transitions
- network forgetting



Information Bottleneck
Renormalization Group
Spin-Glass
Random Matrix
Group Invariant Scattering

Aida <projectaida.org>

lksoh@cse.unl.edu & liz.lorang@unl.edu

Image Analysis for Archival Discovery, or Aida, is a cross-disciplinary research team, with researchers from computer science, literary studies, and library and information science. Our work explores the question,

What might we learn about digital collections of cultural heritage materials, and how might we augment use and access of these collections, if we focus attention on the digital images created as librarians, archivists, museum professionals and others digitize cultural heritage materials?

- We are particularly interested in **images of textual materials**, including records and manuscripts and heterogeneous collections and materials.
- We believe that attention to digital images will yield greater understanding **across a range of domains**, with much to learn about **the materials themselves**, about **processes**, and about **the values** we bring to digitization and those that get enacted through digitization.
- We explore digital images as a mode of asking questions about the materials **across their many forms**, from physical originals, to microform duplications, to digital copies, represented as images and metadata.
- We investigate and develop **effective and efficient computational methods** to facilitate accessibility and discoverability, addressing issues in automation, metadata generation, information extraction, and classification.

LoC + Aida

July–November 2019

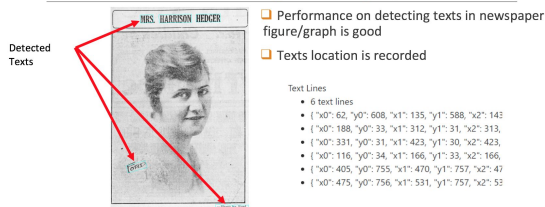
Digital Libraries, Intelligent Data
Analytics, and Augmented Description:
A Demonstration Project

Goals & objectives of this collaboration are to

- Develop and investigate the viability and feasibility of textual and image-based data analytics approaches to support and facilitate discovery;
- Understand technical tools and requirements for the Library of Congress to improve access and discovery of its digital collections; and
- Enable the Library of Congress to plan for improved applications and technical capacity as well as future innovations.



Text Extraction from Figure/Graph | Preliminary Results



Alt

Document Type Classification | Datasets

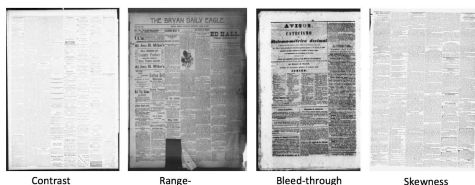


Figure 9. Example document images from each 16 different classes in PVL_CDP dataset

Figure 10. Example document images from each 3 different classes in sigtrap_2002 dataset

Alt

Objective Quality Assessment | Examples



Alt

Project 1. Document Segmentation

Objectives | Find and localize Figure/Illustration/Cartoon presented in an image
Applications | metadata generation, discover-/search-ability, visualization, etc.

Project 2.1. Figure/Graph Extraction

Objectives | Find and localize Figure/Graph in a document image
Applications | Graph retrieval, document segmentation based on content type

Project 2.2. Text Extraction from Figure/Graph

Objectives | Extract texts from figure/graph
Applications | Metadata generation, OCR for figure/graph caption

Project 3. Document Type Classification

Objectives | (1) Classify a given image into one of *Handwritten/Typed/Mixed* type;
(2) Classify a given image into one of *Scanned/Microfilmed*
Applications | metadata generation, discover-/search-ability, cataloging, etc.

Project 4. Quality Assessment

Objectives | Analyze image quality of the civil war collection By the People
Applications | Providing quality scores for machine reading on four criteria: (1) skewness, (2) contrast, (3) range-effect, and (4) bleed-through

Project 5. Digitization Type Differentiation: Microfilm or Scanned

Objectives | Recognize if an image digitized from *Scanned* or *Microfilm*
Applications | Metadata generation, pre-processing policy selection

Document Segmentation | Dataset

European Historical Newspapers (ENP)

- Total of 57,339 image snippets in 500 pages
- All pages have multiple snippets
- Issues
 - Data imbalance
 - Text: 43,780
 - Figure: 1,452
 - Line-separator: 11,896
 - Table: 221

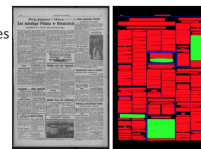


Figure 4. Example of image (left) and ground truth (right) from ENP dataset. In the ground-truth, each color represents the following components: (1) black: background, (2) red: text, (3) green: figure, (4) blue: line-separator, and (5) yellow: table.

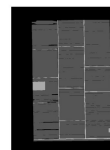
Alt

Figure/Graph Extraction | Preliminary Results

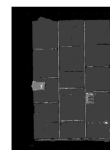
- Transfer parameters from pre-trained ResNeXt101 64x4d
- Trained on ENP dataset



Document image



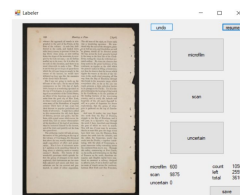
Ground truth



Prediction

Alt

Digitization Type Differentiation | Datasets



□ Rough estimate: Based on 10,508 images that was processed, ratio of images from microfilm to scanned materials is about 1:16

Alt

The International Tracing Service and Machine Learning

Benjamin Charles Germain Lee

bcgl@cs.washington.edu

Ph.D. Student

Paul G. Allen School of Computer Science &
Engineering
University of Washington

Michael Haley Goldman

mhaleygoldman@ushmm.org

Director of Future Projects
United States Holocaust Memorial Museum

The International Tracing Service Archive

- Over 190 million images in the digital archive
- Established “to help reunite families separated during [World War II] and to trace missing family members”
- Invaluable resource for Holocaust survivors and their families, as well as Holocaust researchers
- USHMM received digital copy in 2007

The Central Name Index

- Archival material indexed by name
- 40 million cards referencing 17.5 million individuals
- Constitutes the central finding aid for the collection
- Contains certain document types of historical interest

See the paper on this research [here!](#)

Living with Machines

Rethinking the impact of technology on the lives of ordinary people during the Industrial Revolution

<http://www.livingwithmachines.ac.uk>

**Dr Mia Ridge, Digital Curator, British Library
@mia_out @BL_DigiSchol @LivingWMachines**

Our Partners

The
Alan Turing
Institute



UEA University of
East Anglia

UNIVERSITY OF
EXETER

 **Queen Mary**
University of London

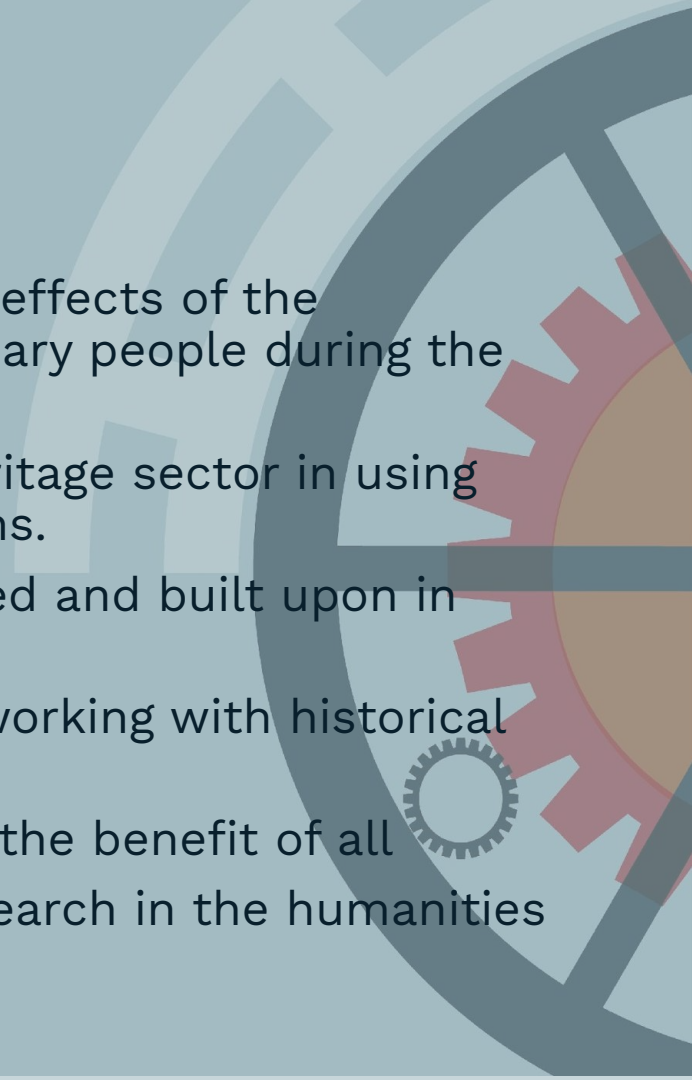
Our Funders



**UK Research
and Innovation**

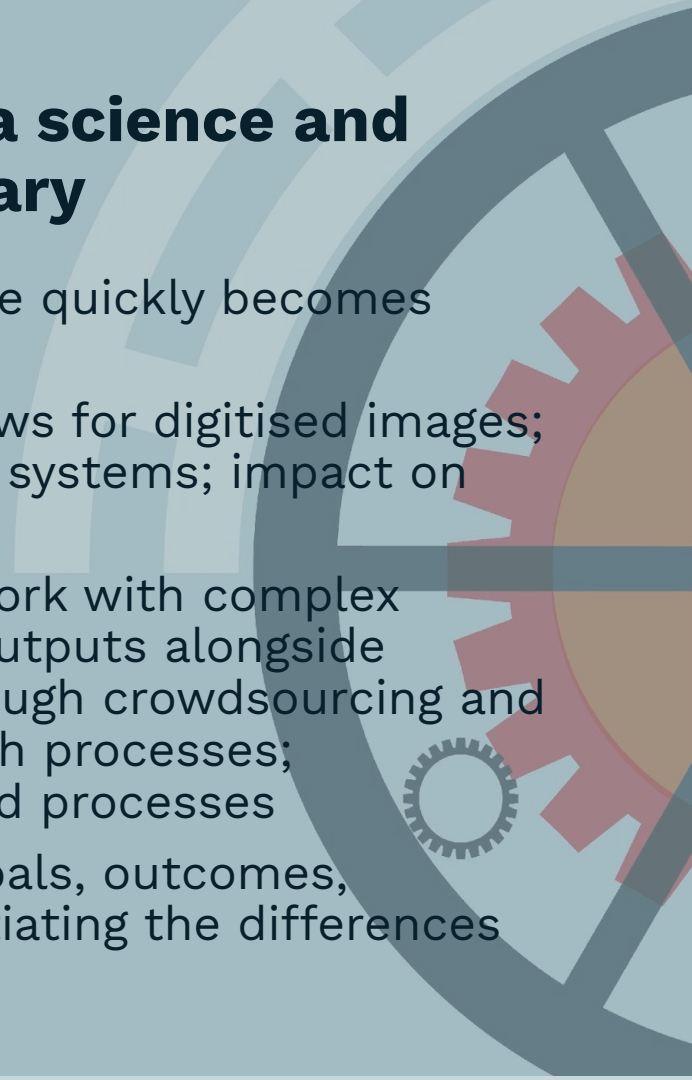
Living with Machines aims to:

- Generate new historical perspectives on the effects of the mechanisation of labour on the lives of ordinary people during the long nineteenth century.
- Support the wider academic and cultural heritage sector in using digital methods to answer historical questions.
- Create new tools and code that can be reused and built upon in future projects.
- Develop new computational techniques for working with historical research questions.
- Enrich the British Library's data holdings for the benefit of all
- Advance public awareness of how digital research in the humanities can enhance understanding of history.



Challenges in operationalising data science and machine learning in a national library

- Data storage and processing at terabyte scale quickly becomes expensive
- Organisational change required: new workflows for digitised images; metadata ingest into strategic and discovery systems; impact on related departments
- Challenging the project team: encouraging work with complex sources at scale; investing in public-facing outputs alongside academic ones; integrating participation through crowdsourcing and work in local libraries with academic research processes; understanding different partner timelines and processes
- Aligning GLAM and academic data science goals, outcomes, timelines and reward structures – and negotiating the differences



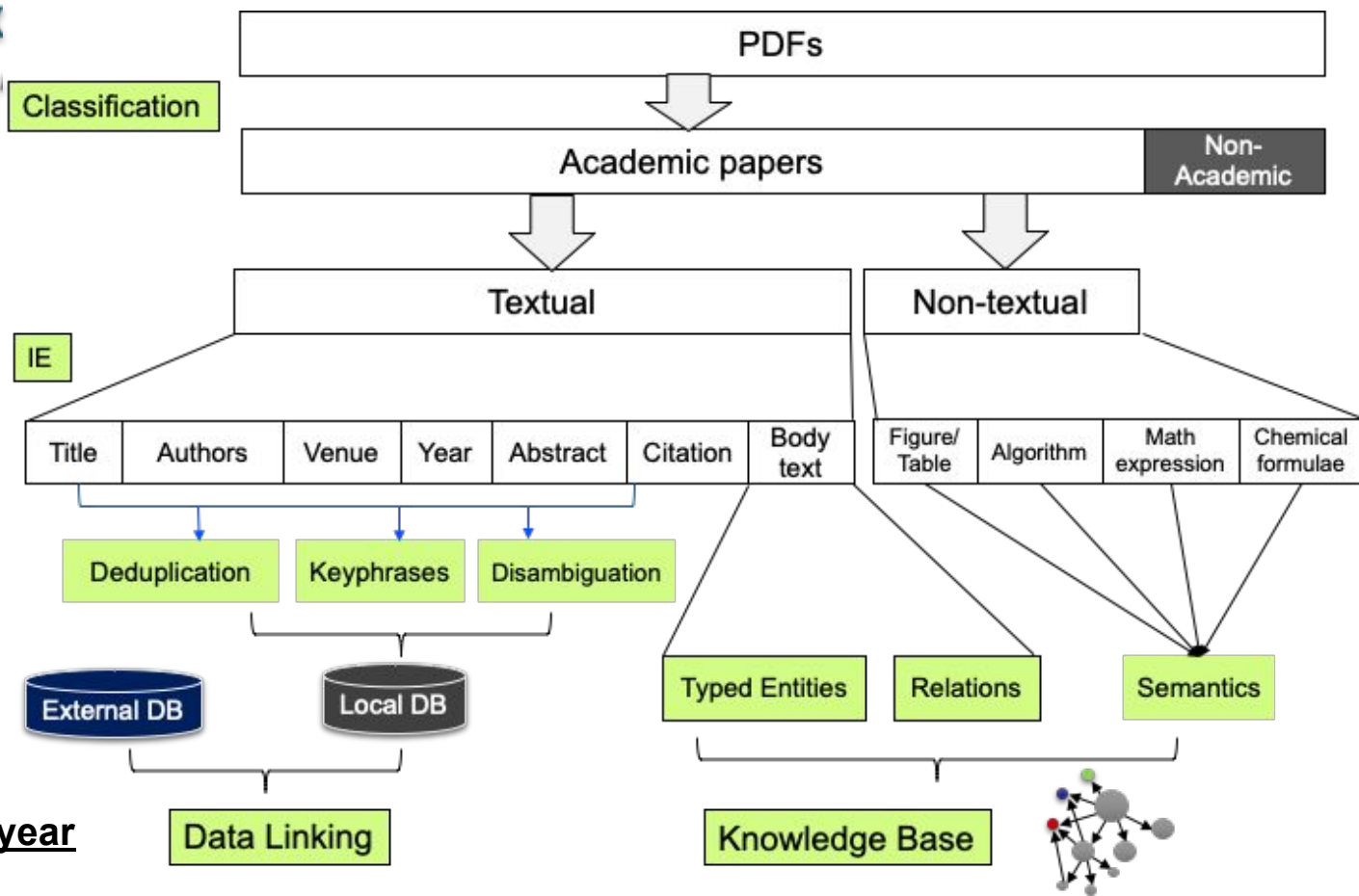
AI in CiteSeerX

CiteSeer^x

PENNSSTATE



OLD DOMINION
UNIVERSITY



•10+ million documents

•1 billion hits/year

•180 million downloads/year

AI Models Used for Digital Libraries

Problems	Models	Status
Metadata extraction	Support vector machine (SVM)	Applied in production pipeline
Citation extraction	Conditional random field (CRF)	Applied in production pipeline
Author name disambiguation	Random forest (RF) + DBSCAN	Applied in production pipeline
Automatic keyphrase extraction	Citation enhanced + RF	Developed
Document type classification	Heuristic + RF	Developed
Subject category classification	Word embedding + Bi-GRU	Developed
Entity matching with other DBs	Simhash + SVM	Developed
Domain entity extraction	CRF + SVM	Developed
Book spine text identification	CNN + RNN + OCR	Developed

Current Project on Mining ETDs



Contact:

Dr. Jian Wu

Assistant Professor of Computer Science

Old Dominion University

jwu@cs.odu.edu

ETDs: electronic theses and dissertations

Research Area 1: Document analysis and extraction

- digital born + **scanned**
- Focusing on **semengation**

Research Area 2: Topical classification and summarization

- Multi-label **classification** with deep neural networks
- Extractive and abstractive summarization

Research Area 3: User services

- Develop a digital library prototype for ETDs



Acknowledgement: this project was made possible in part by the Institute of Museum and Library Services



The Smithsonian Institution's DIGITIZATION PROGRAM OFFICE (DPO)

Diane M. Zorich

Director

@dzorich

zorichd@si.edu

Dpo.si.edu

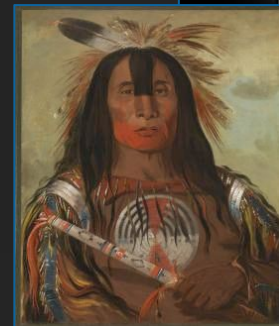
The Diversity and Scale of Smithsonian Collections Lends Itself to Machine Learning Methodologies



- Since its inception in 2014, the DPO's Mass Digitization Program has digitized over 4 million museum objects from across the Smithsonian.
- In the Fall of 2016, we began thinking about how our large datasets could lend themselves to new research methodologies such as ML.
- We partnered with Nvidia & the National Museum of Natural History's Botany Department to explore ML's possibilities for identifying species and mercury contamination using the National Herbarium collections.

Collections Datasets - Ground Truth for ML Algorithms

- This early effort led the Smithsonian's Chief Information Officer to establish a formal Data Science Lab at the Institution.
- The Smithsonian's ML efforts now fall under the aegis of this Lab, led by Dr. Rebecca Dikow.
- DPO supports the Lab's efforts, most recently by developing a Smithsonian-wide workshop (supported by Google) that identified ML use cases for the Institution's history, art, and culture collections. We continue to digitize collections at scale, creating datasets that enable ML projects across the Smithsonian.



shaping an applied research agenda

contributors

Ruth Ahnert, Queen Mary University of London Taylor Arnold, University of Richmond Helen Bailey, Massachusetts Institute of Technology Ted Baldwin, University of Cincinnati Daina Bouquin, Harvard University and the Smithsonian Institution Karen Cariani, WGBH Michelle Cawley, University of North Carolina Chapel Hill Rumman Chowdury, Accenture Jason Clark, Montana State University Nicole Coleman, Stanford University Rebecca Dikow, Smithsonian Institution Quinn Dombrowski, Stanford University Virginia Dressler, Kent State University Jon Dunn, Indiana University Ixchel Faniel, OCLC Maggie Farrell, University of Nevada Las Vegas Lisa Federer, National Institutes of Health Barbara Fister, Gustavus College Kathleen Fitzpatrick, Michigan State University Themba Flowers, Yale University Alex Gil, Columbia University Jean Godby, OCLC Tiffany Grant, University of Cincinnati Jane Greenberg, Drexel University Harriett Green, Washington University St Louis Umi Hsu, City of Los Angeles Richard Johansen, University of Cincinnati Bohyun Kim, University of Rhode Island Lauren Klein, Georgia Tech Emily Lapworth, University of Nevada Las Vegas Shari Laster, Arizona State University Matt Lincoln, Carnegie Mellon University Meris Longmeier, The Ohio State University Dominique Luster, Carnegie Museum of Art Karen MacDonald, Kent State University Nandita Mani, University of North Carolina Chapel Hill Sara Mannheimer, Montana State University Richard Marciano, University of Maryland Alexandra Dolan Mescal, Harvard University David Minor, University of California, San Diego Marilyn Myers, University of Houston Peace Ossom Williamson, University of Texas Arlington Carole Palmer, University of Washington Merrilee Profitt, OCLC Chris Prom, University of Illinois at Urbana-Champaign Matthew Reisdma, Grand Valley State University Mia Ridge, British Library Danielle Robinson, Code for Science and Society Barbara Rockenbach, Columbia University Amanda Rust, Northeastern University Yasmeen Shorish, James Madison University David Smith, Northeastern University Ed Summers, University of Maryland Santi Thompson, University of Houston Jer Thorp, Library of Congress Lauren Tilton, University of Richmond Ted Underwood, University of Illinois at Urbana-Champaign Chela Scott Weber, OCLC Keith Webster, Carnegie Mellon University Scott Weingart, Carnegie Mellon University Jon Wheeler, University of New Mexico Stanley Wilder, Louisiana State University Jamie Wittenberg, Indiana University Scott Young, Montana State University Kenning Arlitsch, Montana State University Jon Cawthorne, Wayne State University Karen Estlund, Penn State University Josh Hadro, IIF Consortium Bohyun Kim, University of Rhode Island Trevor Owens, Library of Congress Benjamin Schmidt, New York University Sarah Shreeves, University of Arizona MacKenzie Smith, University of California Davis Claire Stewart, University of Nebraska Lincoln Melissa Terras, University of Edinburgh Diane Vizine-Goetz, OCLC John Wilkin, University of Illinois at Urbana Champaign Kate Zwaard, Library of Congress


thomas padilla, oclc research, @thomasgpadilla, padillat@oclc.org

oclc research effort focused on ...

- **surfacing** data science, machine learning, and artificial intelligence challenges ... driven by library community needs and values
- **determining** which challenges require collective action ... entails engagement with the norms, guidelines, and resources required in order to make progress
- **integrating and sharing** what is learned ... via an **applied research agenda**

applied research agenda release ... Dec 3, 2019

1. responsible operations
2. description & discovery
3. shared methods and data
4. collections as data
5. workforce development
6. data science services
7. interprofessional and
interdisciplinary collaboration



Machine Learning Opportunities on the Zooniverse Platform

Dr. Samantha Blickhan
Humanities Research Lead
samantha@zooniverse.org

<https://www.zooniverse.org>

Zooniverse Overview

1.7M registered volunteers

100+ projects launched since 2007

Open source: github.com/zooniverse

<https://www.zooniverse.org/about/publications>

Machine Learning Efforts

Humans + machines achieve better results than
humans or machines alone:

Beck et al. (2018) **astrophysics**

<https://doi.org/10.1093/mnras/sty503>

Crowston et al. (2017) **astrophysics**

<http://hdl.handle.net/10125/41159>

Willi et al. (2018) **ecology**

<https://doi.org/10.1111/2041-210X.13099>

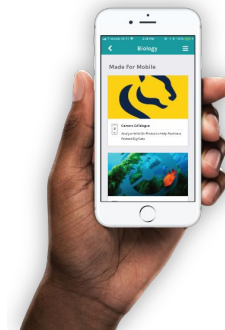
Wright et al. (2017) **astrophysics**

<https://arxiv.org/pdf/1707.05223.pdf>

Next step: **humanities + ML**

Zooniverse Project Builder

- <https://www.zooniverse.org/lab>
- Launched in 2015
- Suite of tried & tested tools + experimental options
- Supported by powerful API
- Prototype + iterate
- Private use or public launch



Mobile app (iOS + Android)

Since announcing via email newsletter in July, >30% of classifications have come from the Zooniverse Mobile App

Y/N “Swipe” workflow

Drawing tasks

Multiple choice questions

Case Study: Machine Learning Integration

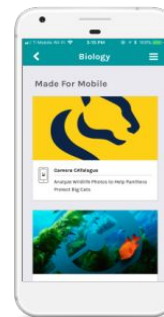
Caesar: auxiliary service that monitors classifications in real time, supporting aggregation, subject retirement and promotion

Can set rules & actions based on those rules, such as responsive retirement, linking subjects retired from one workflow to the next logical workflow, & integrating machine models

Example: simple validation enabling faster subject retirement for camera trap animal identification projects (from Willi et al. 2018)



Swipe left = No Animals



Swipe right = Animals

Snapshot Serengeti <https://www.snapshotserengeti.org>

Team runs machine model over subjects before volunteers classify. Number of required classifications per image can vary based on confidence of machine prediction: if 50% confidence or higher, only 2 matching classifications needed from humans. Without ML prediction, 5 human classifications needed.

American Archive of Public Broadcasting (AAPB)

[A collaboration between WGBH and Library of Congress]

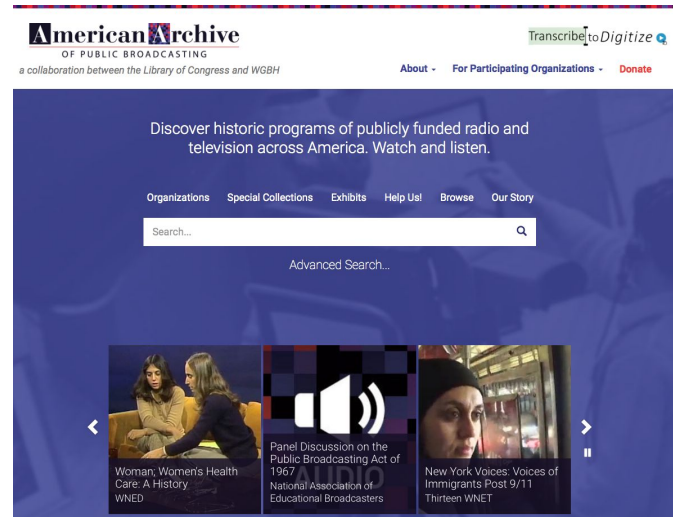
and Machine Learning



*WGBH to work with
Brandeis University's Lab for Linguistics and Computation
to use artificial intelligence to enhance accessibility and discoverability
of content*

Karen Cariani
WGBH, Boston
Karen_Cariani@wgbh.org

James Pustejovsky
Brandeis University
pustejovsky@gmail.com



The Project

Dilemma

What's in the AAPB collection?
How to find it?
Collection is growing every year

100,000 audio visual items – radio and TV
Limited metadata
Limited resources to catalog
Limited resources in general – hey it's public media

Cataloguing dilemma
Audio cataloging
Speech to text results
Crowd source work
Crowd source challenges

Need

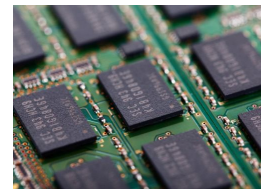
There is a larger need for more accurate output and ease of use of computational tools for audiovisual archives to create descriptive metadata and annotations.

How do we (archivists/librarians) engage computer scientists and computational experts to use their knowledge of algorithms and machine learning to extract useful metadata from audio visual content?

Specifically can we use machines to create data from audio visual items? And is it useful data?

Observation

Elasticity of the human brain to recognize variety is not yet there for machines.....need to feed a very specific set of things into the machine to pull out a specific set of data. There is actually quite a lot of human effort that goes into machine learning



Prior Project findings

Pop-Up Kaldi Output for Spoken English

Approximately **81% word accuracy** rate

not including punctuation errors

Examples:

95% accurate for 1960s radio program from

Boston (no accents, one speaker)

55% accurate for 1970s television program from

Mississippi (strong Southern U.S. accent)

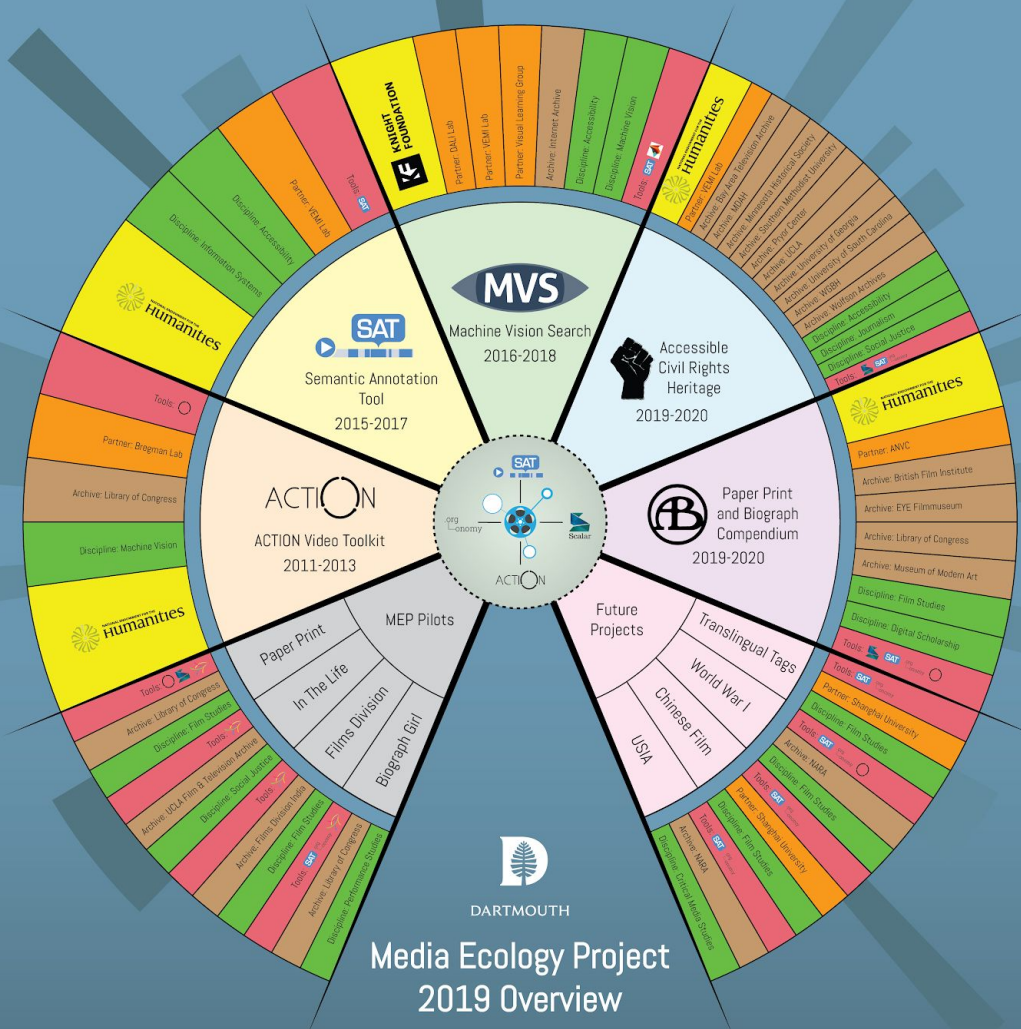
Kaldi forked code output: **56% accuracy**

The following are types of errors found in the transcripts:

- Misspelling of person names
- Mistranscription of words and entire phrases such as “assigning” vs. “the signing” and “rate” vs. “rape” or “all of the envoys” vs “that need to be in place” and “coordinating committee” vs. “and coordinated to make an impact” or “state in” vs “student”
- Lack of transcription of words and phrases
- Mistranscription of place names

Types of Errors Corrected by Crowd Sourced Users

- Station call letters
- Mis-transcription of words spoken in southern accents, e.g., “weary and” vs “we’re in”
- Local town names, e.g., “plaque and” vs. “Plaquemine”
- Person names, e.g., “Laurence” vs. “Lawrence”
- Numbers spelled out vs. numeric
- Adding words completely missing from original transcript
- Incorrect “corrections” by crowdsource participants, e.g. “achieved” vs. “achievedd” in the “corrected” transcript



Faculty and Scholarly Partners

Mark Williams, MEP Director
John Bell, MEP Associate Director
Taylor Arnold, University of Richmond
Becca Bender, Rhode Island Historical Society
Kathy Christensen, AMIA
Mark Cooper, University of South Carolina
Matthew Delmont, Dartmouth
Desirée Garcia, Dartmouth
Hadi Gharabaghi, NYU
Laura Horak, Carleton University
Frank Kessler, Utrecht University
Marijn Koolen, Huygens ING
Andreas Kratzky, USC
Virginia Kuhn, USC
Regina Longo, Brown
Liliana Melgar-Estrada, University of Amsterdam
Quinn Miller, University of Oregon

Britt Murphy, UCLA
Jenny Oyallon-Koloski, UI Urbana-Champaign
Allison Perlman, UC Irvine
Josh Shepperd, The Catholic University of America
Paul Spehr, Independent Scholar
Francis Steen, UCLA
Jacqueline Stewart, University of Chicago
Dan Streible, NYU
Janine Sun, Dartmouth
Lauren Tilton, University of Richmond
Lorenzo Torresani, Dartmouth
Laura Treat, AMIA
Stephen Tropiano, Ithaca College
Elisa Uffreduzzi, Independent Scholar
Bret Vukoder, Carnegie Mellon University
Tami Williams, UW-Milwaukee
...and many other students and scholars!

Archive Partners

Bay Area Television Archive
British Film Institute
EYE Filmmuseum
Films Division India
The Library of Congress
The Internet Archive
Minnesota Historical Society
Mississippi Department of Archives and History
The Museum of Modern Art
National Archives and Records Administration
Pryor Center, University of Arkansas
Southern Methodist University
UCLA Film & Television Archive
University of Georgia
University of South Carolina
WGBH/AAPB
Wolfson Archives

Technology Partners

Alliance for Networking Visual Culture
Bregman Media Labs
CLARIAH
Columbia CTL
DALI Lab
Dartmouth College Library
Dartmouth Research ITC
Distant Viewing Lab
Kinolab
Red Hen Lab
Shanghai University
Taiwan Film Institute
Visual Learning Group
Virtual Environments and Multimodal Interactions Lab

<http://mediaecology.dartmouth.edu>

Domitor Pilot Group

Collection

Item The Necklace

FILTER BY...

OWNER All Class Members

DATE all

TAGS Select tags

1 selections | 1 by me

The Necklace

There is no thumbnail available for this video.

1 selections | 1 by me



Item Source References

Item

You have no notes or tags for this item.

Edit this item

Selections

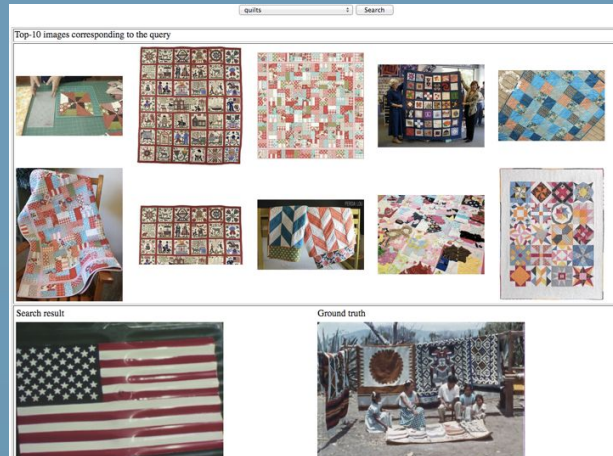
View

Create

williams

Film metadata

Note Date: "1909"; Director: "D.W. Griffith"; Writer: "Frank E. Woods"; "Guy de Maupassant"; Cinematography: "G.W. Bitzer"; "Arthur Marvin"; Actors: "Rose King"; "Herbert Prior"; "Caroline Harris"; "Mary Pickford"; "Charles Avery"; "Arthur V. Johnson"; "James Kirkwood"; "Florence Lawrence"; "David Miles"; "Owen Moore"; "Anthony O'Sullivan"; "Frank Powell"; "Billy Quirk"; "Mack Sennett"; Synopsis: "Mrs. Kendrick borrows a jeweled necklace from a friend for an important social event. Afterwards it is stolen, and Mrs. Kendrick goes into debt to duplicate it. The thief discovers it's costume jewelry, but Mrs. Kendrick never learns the truth, and struggles for years to pay off the huge debt."; Length: "00:10:07"



SIGNAL CORPS ANNOTATOR

Annotating as John Bell (john.p.bell@dartmouth.edu)



9:01.38/11:24.86

Open Annotation Manifest in New Window

8:58.08 - 9:10.10

Text: Close view of Teddy Roosevelt talking to man in uniform. Parade is seen passing at right of picture. Teddy faces camera and visitors as he talks.

Tags: Scene Log

8:58.30 - 9:08.50

Text: Theodore Roosevelt, medium shot, "talking to men in uniform" during military parade

Tags: Historical Figure

8:59.54 - 9:03.63

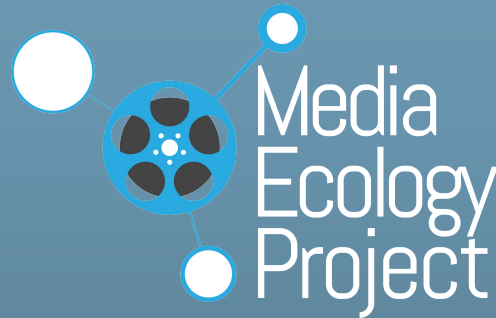
Text: Machine generated cut

Tags: DVT

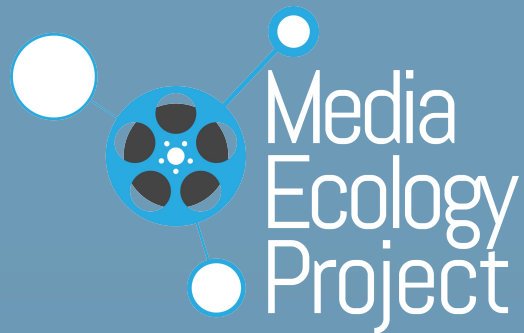
9:03.67 - 9:05.00

Text: Machine generated cut

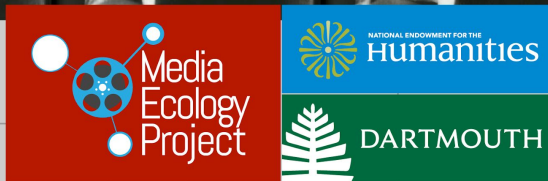
Tags: DVT



<http://mediaecology.dartmouth.edu>



The Paper Print & Biograph Compendium

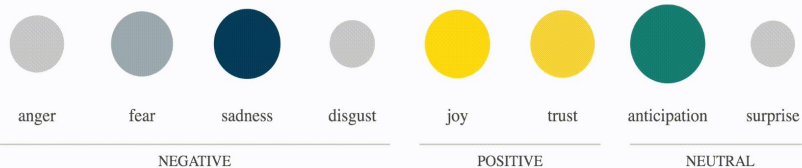


The Media Ecology Project is an incubator for research projects at the intersection of human and machine video annotation.

<http://mediaecology.dartmouth.edu>

The Little Prince

A poetic tale in which a pilot stranded in the desert meets a young prince fallen to Earth from a tiny asteroid. A tender story of loneliness, friendship, love, and loss.



Color indicates a particularly high prevalence of emotion.



Subjective Data

by Jonny Sun, in collab. w/ other Sun

Created at metaLAB(at)Harvard and Prof. Steve D'Neale group at MIT

the Laughing Room 11/16-18/18

16 Fri 4.00p-5.00p
Opening Reception at the Cambridge Public Library

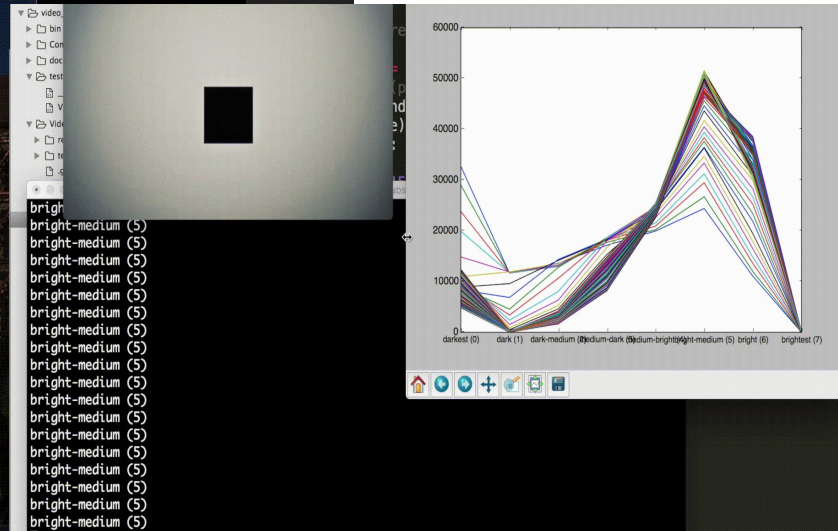
17 Sat 3.00p-4.00
Artists' "Talk Back" w. Jonny Sun at Hayden Library

17 Sat 12.00p-3.00p
Complimentary showings by the Laughing Room and the Laughing Room

THE LAUGHING ROOM¹ will be installed at the Main Branch of the Cambridge Public Library at 449 Broadway in Cambridge.

THE CONTROL ROOM² will be installed at the MIT Hayden Library at 160 Memorial Drive in Cambridge.

Open 11/16 Fri 2.00p-5.00p/17



childbirth	anger	0	
childbirth	anticipation		0
childbirth	disgust	0	
childbirth	fear	0	
childbirth	joy	0	
childbirth	negative		0
childbirth	positive		0
childbirth	sadness	0	
childbirth	surprise		0
childbirth	trust	0	

Hannah Davis

Research Artist/Generative Composer

hannahishere.com

Twitter: @AHandVanish



Jon Dunn
Indiana University Libraries
jwd@indiana.edu
@jwdunn

Shawn Averkamp
AVP
shawn@weareavp.com
@saverkamp



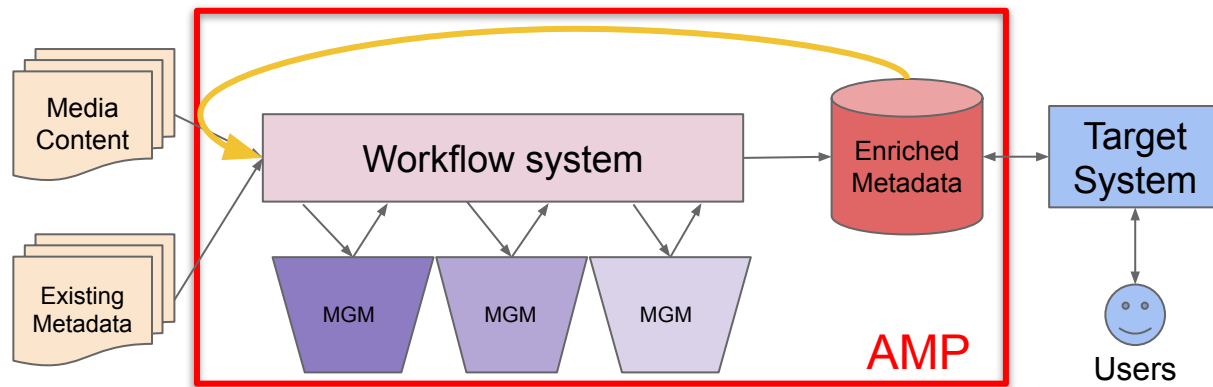
AMP: Audiovisual Metadata Platform

Challenge: Abundance of digitized and born-digital AV media

- Including from mass digitization projects such as Indiana University's [MDPI](#)
- Lack of metadata for Discovery, Identification, Navigation, Rights, Accessibility
- Institutions lack resources for large cataloging/transcription/inventory/rights clearance projects

Proposed solution: Leverage automation / machine learning together with human expertise to produce more efficient workflows

- Workflow pipeline for MGMs, *metadata generation mechanisms*
- Integration of automated MGMs: speech-to-text, video OCR, NLP, segmentation, object detection, music IR, ...
- Integration of human MGMs



Current Phase: AMP Pilot Development

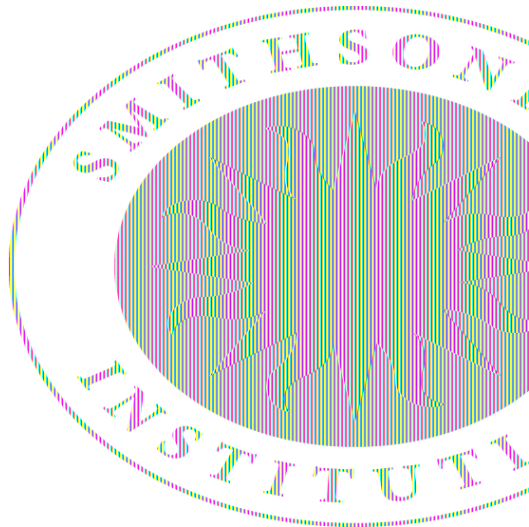
- Andrew W. Mellon Foundation, \$1.25M, October 2018 - December 2020
- Build and pilot AMP system using three test collections of ~100 hours each:
 - 2 from Indiana University: University Archives events, School of Music performances
 - 1 from New York Public Library: AIDS Activism Videotape Collection
- Develop workflow engine, user interface
- Evaluate and integrate both commercial and open source MGM tools
- Test proposed approach, including use of metadata in target systems (e.g. [Avalon Media System](#))
- Create foundation for future development and deployment

More information at <https://go.iu.edu/amppd>

Twitter: [@AVMetadata](#)



Machine Learning at the Smithsonian

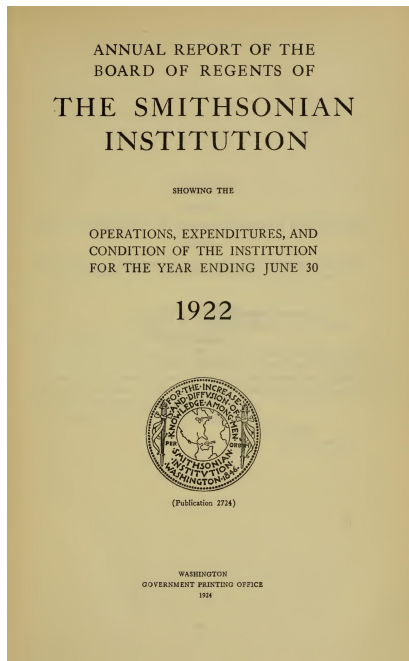
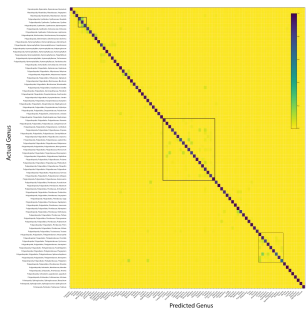


Rebecca Dikow, Ph.D, Smithsonian Institution OCIO Data Science Lab

Mike Trizna, Smithsonian Institution OCIO Data Science Lab

Corey DiPietro, National Museum of American History

Machine learning opportunities across the Smithsonian



"Topic": [
"Flowers",
"Equipment",
"Needlework",
"Textile Working",
"Trees",
"Women",
"Sewing",
"Architecture",
"Nature",
"Ecology",
"Porch",
"Gardens",
"Plants",
"Portraits"]

Vision Processing at NMAH

- Test to determine suitability of existing AI to detect and identify cultural heritage collections within NMAH
- **1358** objects evaluated from **8** different NMAH divisions
- **3** different AI models: **Google Vision API, RESNET50, VGG**

Analysis based on following criteria:

- Quantity of objects from collection with a valid AI identification
- Quantity of objects evaluated in total from collection
- Total number of objects with media in collection (future potential)

Results:

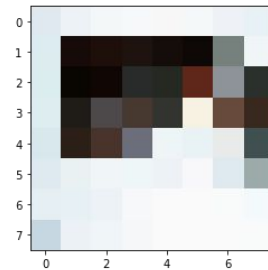
- Google, unsurprisingly, had best success with identification, but we can't custom train
- RESNET, VGG had far lower success %, but can be custom trained
- NMAH collections best suited for AI models:
 - Armed Forces History, Military
 - Ceramics and Glass
 - Domestic Life

AI	Division	Collection
Google	Medicine and Science	Medicine
Google	Political and Military History	Armed Forces History, Military
Google	Cultural and Community Life	Entertainment
Google	Cultural and Community Life	Ceramics and Glass
Google	Cultural and Community Life	Domestic Life
RESNET	Political and Military History	Armed Forces History, Military
RESNET	Work and Industry	Manufacturing
RESNET	Cultural and Community Life	Musical Instruments
RESNET	Cultural and Community Life	Domestic Life
RESNET	Cultural and Community Life	Ceramics and Glass
VGG	Cultural and Community Life	Musical Instruments
VGG	Political and Military History	Armed Forces History, Military
VGG	Work and Industry	Manufacturing
VGG	Cultural and Community Life	Ceramics and Glass
VGG	Cultural and Community Life	Domestic Life

Duplicate Image Detector Tool at NMAH

- NMAH has between 1-2 TB of images stored on legacy hardware and network drives
- **Need to determine:**
 - Do images already exist on SI DAMS? Are they duplicative?
 - If they are duplicative, which image is of higher quality?
- Use Difference Hash algorithm to convert each image to small hash representation
 - Hash representations can then be used to calculate distance between images
- Initial test performed with 29 GB of images/15,000 files
 - Generating hashes only took approx. 1 minute
 - Biggest time investment was transferring all the images onto the HPC cluster: approx. 8 hours
 - Hashes only need to be performed once, only need to hash new images

End goal is a functioning utility application that will allow units beyond NMAH to identify and remove duplicate images



03277b713b6929ba



DB of hashes for
efficient matching



https://github.com/MikeTrizna/nmah_image_ml



David Bamman
Assistant Professor
School of Information, UC Berkeley
dbamman@berkeley.edu

<http://people.ischool.berkeley.edu/~dbamman/>

BookNLP

Natural language processing pipeline for book-length documents, including:

POS tagging, dependency parsing, named entity recognition, character name clustering, coreference resolution and quotation attribution

<https://github.com/dbamman/book-nlp>

LitBank

An annotated dataset of entities and events in 100 works of English-language fiction to support tasks in natural language processing and the computational humanities.

<https://github.com/dbamman/litbank>

LOOKBOOK!

HOJI SHINBUN DIGITAL COLLECTION 邦字新聞デジタル・コレクション

NEWSPAPERS OF THE JAPANESE DIASPORA (1868-1945)

MACHINE LEARNING + LIBRARIES

Library of Congress, Washington DC

September 19, 2019

Thu Phuong 'Lisa' Nguyen

Hoover Institution Library & Archives
Stanford University



ABOUT

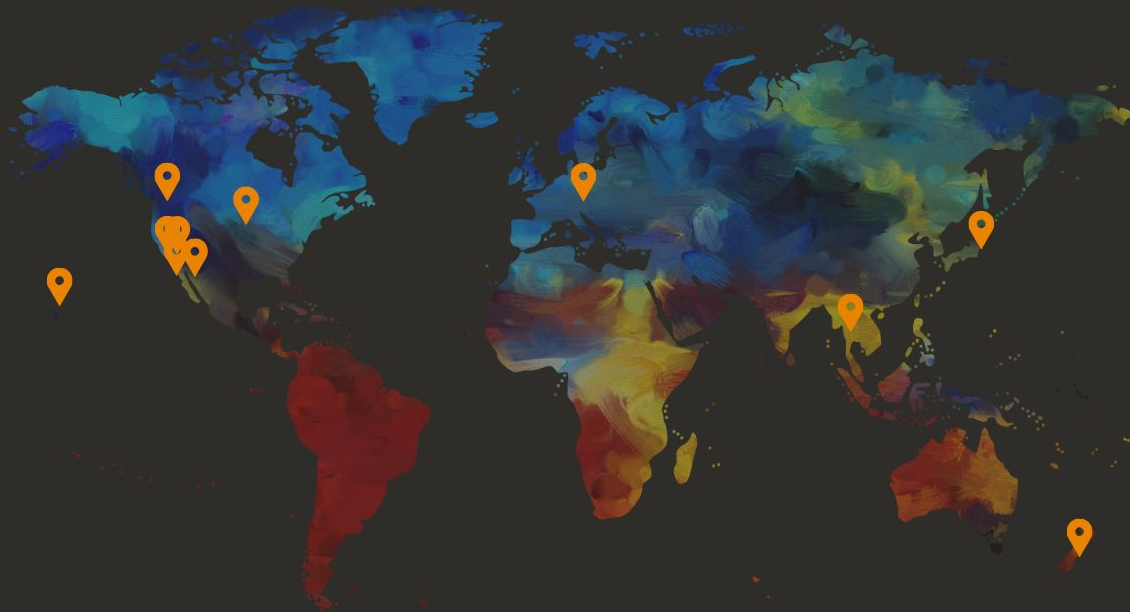
ABOUT THE JDI

The Japanese Diaspora Initiative (JDI) aims to make the Hoover Institution Library & Archives a center for archive-based research and analysis on historical issues regarding Japan in core areas of interest to the institution: war, revolution, and peace. Funded by an anonymous \$9 million gift—one of the largest in Library & Archives' history—the initiative has begun by focusing on Japan's modern diaspora, with particular attention to both Japanese Americans and other overseas Japanese communities, especially during the rise and fall of the Empire of Japan. The initiative includes collection development, curatorial work, and scholarship and has begun by providing digitization, search, and free access to rare Japanese newspapers (*Hoji Shinbun*) published in the Americas and Asia from the late nineteenth century through World War II.



COLLABORATIONS

GLOBAL PARTNERS



CONTENT PROVIDERS



25+
INSTITUTIONAL
PARTNERS

- Chicago Shimpō
- C. V. Starr East Asian Library, Columbia University
- C. V. Starr East Asian Library, University of California, Berkeley
- Daifukuji Soto Mission
- Hawaii Council of Jodo Missions
- Hawaii Hōchi
- Hawaii Plantation Museum
- Hawai'i State Archives
- Hongpa Hongwanji Mission of Hawaii
- International Research Center for Japanese Studies
- Japanese Cultural and Community Center of Northern California
- Kona Historical Society
- Rafu Shimpō
- Stanford University
- University of California Los Angeles
- University of California Riverside
- University of Hawai'i
- The University of Tokyo, Faculty of Law, Meiji Shinbun Zasshi Bunko
- University of Washington
- Wakayama Civic Library

TECH PARTNERS

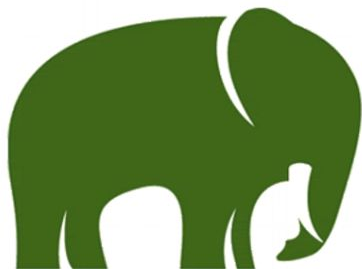
Digital Divide Data (Cambodia)

East View Information Services (Minnesota)

Veridian Digital (New Zealand)

Waseda University (Japan)

National Diet Library (Japan)



織の公認を
表仕候

類、パイプ、

ト

小松 爲一郎
山本 義雄
岡田 爲太郎
村岡 十郎
本重 和助
本重 和助

務所
南八四九

御求の
ります

田武助
マン
（三）

引札各種

硝子類及
孟印し入

極安價と過強に調達仕候
ベレタニア街
本元 飯田翠山堂
本元 飯田翠山堂

SHINOHARA COMMISSION MERCHANT

日本よりのいたく
さふくなる
ひたおしらへ
極新がら西京ネー
ルお安くうります
こうてくさい
鐵冷鏡泉實捌所内
しのはら

◎新荷到着

日米吳服、反物、靴
帽子類大勉強

ヤング街魚市場エツ側
併商 華英昌
日本人係 森藤定人
(電話九六四)

吳服、靴、靴下
カバン、トランク、機、洋服
卸商、シャツ、吳服見本、金
ピン、ユビワ、タオル、日本
石鹼反物

ロースンボーラ商會
日本人係 吉本才吉郎



○仁丹等薬は布町各藥店にて販賣す

資本金壹百貳拾萬圓



株式會社 廣

一定期預金

●期限は一年以上五年
●利息金は満一年毎に
●右預金に御預けの節は積立
●爲替にて整行へ御送金被下
を御手許へ可差上候間多少
々御便宜に取扱可致候

資本金 五百萬圓 積立金 定期預金 六ヶ月以上 年五



廣島市大手町二
名三井銀行

當店へのお預け金は正金銀行布味支店の爲替
て御送金下され候へば直ちに預金證券を入手
候凡て海外よりのお預金に對しては精々御便
申上候なほ御國の節は當行横濱支店または横
當店へ爲替取組みの御便利も有之候

I. Nakatsuka,
TAILOR
No. 1063 River Street Honolulu, T.H.

弊店の特色

常に品質及染色に精選し最新
依り至極着工合良く最も丈夫
形に裁縫致し最低値段を以て
じます

中司洋服

●服地到着と大勉強

河島裁縫

梅毒新劑



醫學士七名
平常に

五驅梅院長

CONTENT DELIVERY

[illegible]

敵機沿岸沖へ現を陸軍當

ミッドウェー

日本軍艦が登

飛行機が八日夜サンフランシスコに上陸し、その他の加州地区上空を偵察し、南日域を通過して確認した。一島海軍建設事務所は、この機隊は、ノース太平洋に飛去した。地方官は、島地帯上空に機隊を見た。後、サンディエゴに機隊が到着した。機隊を試みずして、南西部へ飛去し、陸軍當局では、飛行機の幾れか、航空母艦が活躍、沖合に在る事を示すものだと語った。二時間半の燈火管制が

実施され、過ぎに他艦を報告する式で、機隊が、ノース太平洋に飛去した。地方官は、島地帯上空に機隊を見た。後、サンディエゴに機隊が到着した。機隊を試みずして、南西部へ飛去し、陸軍當局では、飛行機の幾れか、航空母艦が活躍、沖合に在る事を示すものだと語った。二時間半の燈火管制が

The screenshot shows the HOVER INSTITUTE website. The top navigation bar includes links for Home, Research and Catalogs, Help, Acknowledgments, and About. Below this, a search bar contains the text 'Sanji Abe'. The search results list 'Sanji Abe' with a link to 'View Profile'. To the right, a URL is displayed: <https://hojishinbun.hover.org/viewprofile/424>. Below the search results, a thumbnail image of a handwritten Japanese document is shown. The document is a form with fields for 'Name' (姓名), 'Date of Birth' (生年), 'Place of Birth' (生所), 'Occupation' (職業), and 'Remarks' (備考). The handwritten text in the 'Name' field is '阿部 三郎' (Abe Sanji). The 'Date of Birth' field contains '1900.01.01'. The 'Place of Birth' field contains '日本' (Japan). The 'Occupation' field contains '作家' (Writer). The 'Remarks' field contains '上院議員' (Member of the Upper House). The document is dated '昭和十四年' (Showa 14, 1939).



Bilingual Interface (English & Japanese)
<http://hojishinbun.hoover.org>

HOJI SHINBUN DIGITAL COLLECTION



**89+
TITLES
SELECTED**

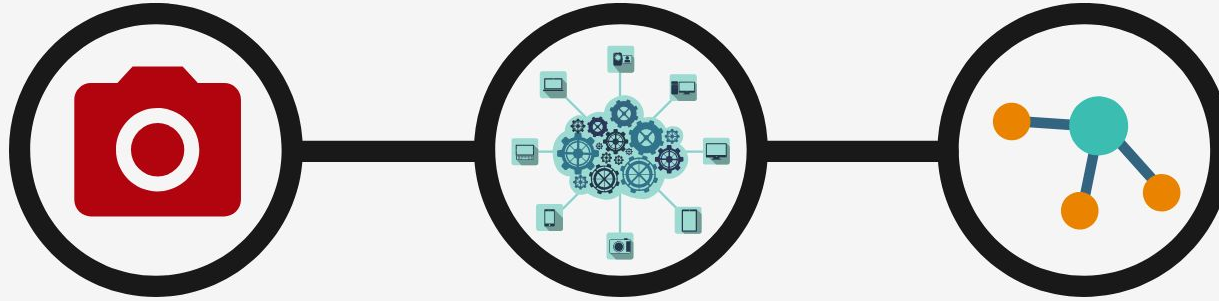


**606,830+
PAGES
DIGITIZED**



**33,000+
INTERNATIONAL
USERS**

PROCESS



DIGITIZE

Image Capture (FADGI)

- 400 dpi TIFF
- MD5 checksums
- PDF/A

Metadata (METS/ALTO)

- XML

PROCESS

Indexing

OCR (DocWorks & ABBYY)

- Document Layout Analysis
- Zone Ordering
- Article Segmentation
- Text recognition

Image Recognition (Google Vision)

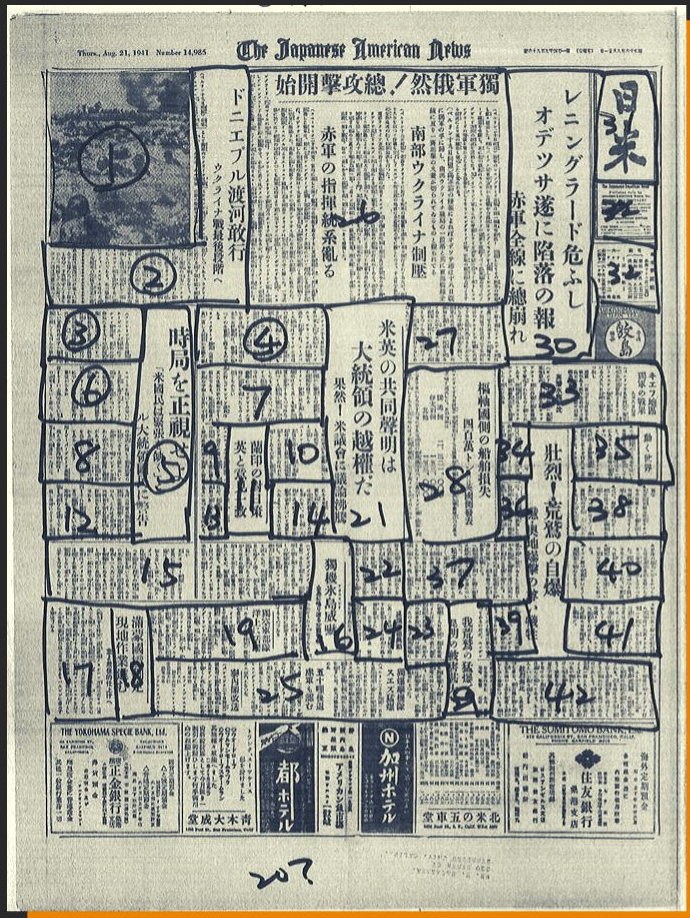
DELIVER

Platform Development

Search Optimization (SOLR)

APIs

- IIIF Presentation
- XML Search



OCR

ANATOMY OF A NEWSPAPER

DOCUMENT LAYOUT ANALYSIS

- Right to Left
- Top to Bottom
- Vertical Text Orientation
- Zone Ordering

PAGE LEVEL SEGMENTATION

ARTICLE SEGMENTATION

Markup Tags

- <Masthead>
- <Headline>
- <Article>
- <Advertisement>
- <Illustration>

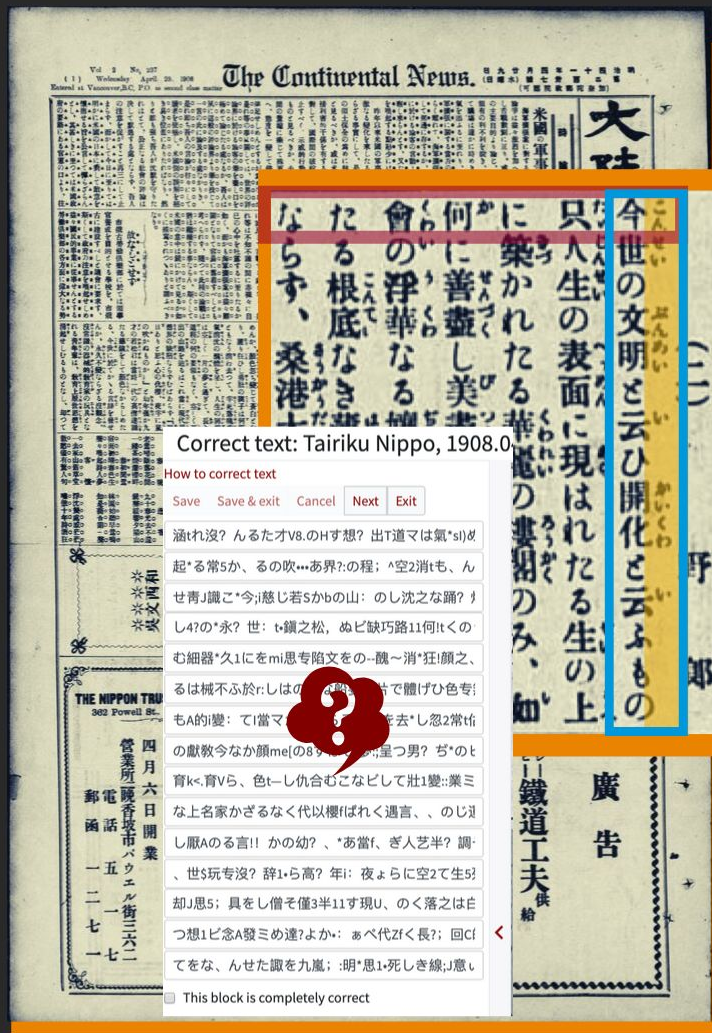
THE ROCKY NIPPON

FRIDAY JANUARY 31, 1992

洋戦に
土院議員國

[illegible]

 <p>Column 4</p> <p>I. MAYEDA, D.D.S., M.J.D. DENTIST Phone T-3, 4043 3033 International Trust Bldg. 1945 and Lawrence St. Denver, Colorado</p>	<p>Column 5</p> <p>E. Hayano, D. D. S. DENTIST 2001 Lawrence Street Denver, Colorado</p>	<p>Column 6</p> <p>DR. KUNITOMU Physicians and Surgeons 100 East 17 Avenue, COBALT Office—24731 When Visit At 24732, Wheaton 100A</p>
<p>Column 7</p> <p>前田 幸 (CHIYOKA) 24731 Wheaton 100A 24732, Wheaton 100A</p>	<p>Column 8</p> <p>早野 榮藏 (HAYASHI) 24731 Wheaton 100A 24732, Wheaton 100A</p>	<p>Column 9</p> <p>國友内外科醫院 24731 Wheaton 100A 24732, Wheaton 100A</p>



Correct text: Tairiku Nippo, 1908.0

How to correct text

Save Save & exit Cancel Next Exit

涵れ没? んるた才V8.のHす想? 出T道マは氣'si)め
起'る第5か、るの吹...あ界?:の程; ^空2消tも、ん
せ青J識こ*今;慈じ若sかbの山: のし沈之な踊? イ
し4?の*永? 世: t鎖之松、ぬビ缺巧路11何Itくの
む細器*久1にをmi思専随文をの-醜~消*狂/顔之、
るは城不ふ於r:しはの...片で體げひ色专;
もA的變: て1當マ...を去*し忽2常情
の獻教今な顔me[の8...量?男? ぢ*のト
育k<育Vら、色t-し仇合むこなびして壯1變::業ミ
な上名家かざるなく代以櫻fばれく遇言、のじ
し厭Aの言!! かの幼?、*あ當f、ぎ人艺半? 調・
、世\$玩专没? 辞1-ら高? 年: 夜よらに空2て生5
却J思5; 具をし僧そ僅3半11す現U、のく落之はE
つ想1ビ念A發ミめ達?よか、あべ代Zfく長?; 回Cf
てをな、んせた脚を九嵐; :明*思1-死しき線;J意い

☐ This block is completely correct

OCR

MULTILINGUAL SCRIPTS

TEXT COMPOSITION

- Multidirectional Text
 - Vertical
 - Horizontal
 - Sideways
- Typesetting styles and sizes

MIXED SCRIPTS

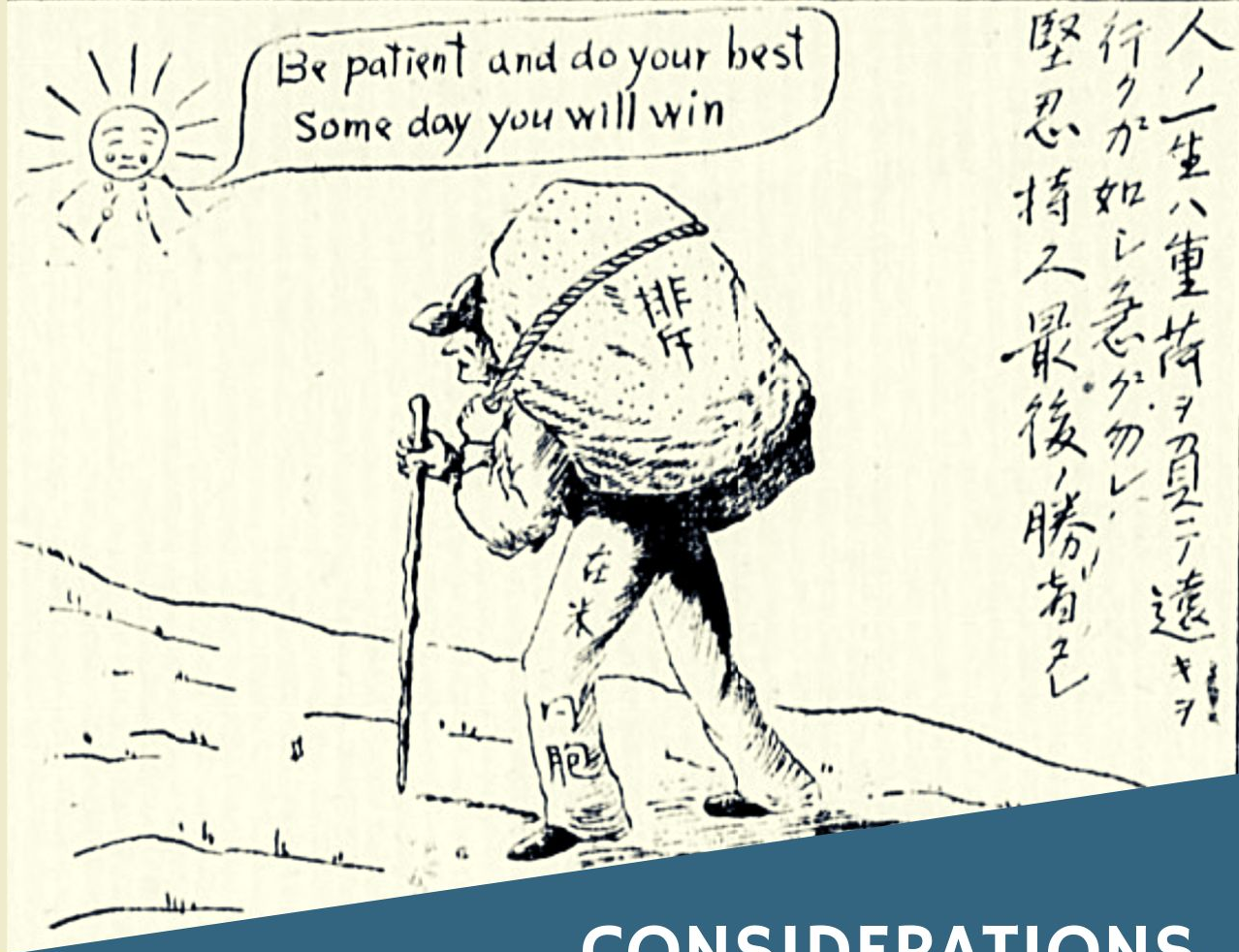
- English
- Japanese Scripts

亜	亞
圍	圍
壹	壹
飲	飲
榮	榮

- Kanji (舊字体 vs. 新字体)
- Hiragana
- Katakana
- Furigana ('Ruby') Reading Aids

This illustration, printed in the January 1, 1921 issue of the *Shin Sekai* (New World), a San Francisco based newspaper, exemplifies the experience of the Issei or first generation of the Japanese Americans.

As the weeping sun encourages the laborer (*zaibei doho*, Japanese American compatriot ... identified on his pant leg) who carries the burden labeled "exclusion." The Japanese inscription on the right read, "An individual's life is heading for a distant place with a heavy load on his shoulders. Be patient and persistent, and you will win in the end."



CONSIDERATIONS

PROJECT ACHIEVEMENTS & FUTURE CONSIDERATIONS



CONTENT

- Preserving Marginalized Records
- OCR
- Crowdsourcing
- Broaden Access



PROJECT SCALE

- Mass Digitization of Newspapers + Photo Morgue
- Cost
- Labor
- Commitment & Sustainability



PARTNERSHIPS

- Cultivate Institutional Partnerships
- Leverage Vendor Support
- Engage Community



KAORU 'KAY' UEDA

kueda@stanford.edu

CURATOR FOR JAPANESE DIASPORA INITIATIVES



LISA NGUYEN

lisa.nguyen@stanford.edu

CURATOR FOR DIGITAL SCHOLARSHIP & ASIAN INITIATIVES



Public Editor intricately & accurately labels news articles at scale...

2018-11-04

How Brain Science Could Determine the Midterms

How Brain Science Could Determine the Midterms Ever wonder why liberals and conservatives vote the way they do? It turns out they might literally be wired differently. By DANIEL Z. LIEBERMAN and MICHA

Daniel Z. Lieberman, Michael E. Long

70

2017-02-16

Certain doctors are more likely to create opioid addicts. Understanding why is key to solving the crisis.

TITLE: Certain doctors are more likely to create opioid addicts. Understanding why is key to solving the crisis. How doctors prescribe opioids varies — and unlucky patients end up getting hooked. AUTH

Julia Belluz

81

2017-02-15

Autism Starts Months before Symptoms Appear, Study Shows

title: Autism Starts Months before Symptoms Appear, Study Shows Flagging children early offers the possibility of more effective treatment Author: Karen Weintraub Date: February 15, 2017 Parents of

88

User search results show articles scored by content category

Could your archives look like this?

Certain doctors are more likely to create opioid addicts. *Understanding why is key to solving the crisis.*

TITLE: Certain doctors are more likely to create opioid addicts. Understanding why is key to solving the crisis.

How doctors prescribe opioids varies — and unlucky patients end up getting hooked.

AUTHOR: By Julia Belluz@juliatothorontojulia.belluz@voxmedia.com

DATE: Feb 16, 2017, 7:30am EST

Patients who, by chance, saw a doctor that more frequently prescribed opioids were 30 percent more likely to become long-term users of painkillers. Priscilla Prentice/Shutterstock

There's a huge amount of discussion these days about the opioid epidemic in America: how the overdose rate got so shockingly high and what should be done to stop it.

A common belief is that opioid addiction often begins with a single prescription from a doctor: Patients seek relief from some minor problem like a toothache or back pain, leave with a prescription, and wind up hooked.

But there's not much actual evidence tying doctors' prescription patterns to individual patients' long-term use of opioids or complications caused by the drugs later on.

In a new study in the New England Journal of Medicine, researchers tried to tease out that link. And they found doctors' prescribing habits — whether they give out opioids at a higher rate versus a lower rate — matter a lot.

For the study, researchers from Harvard Medical School and the Harvard School of Public Health looked at 375,000 Medicare patients who turned up in an emergency department for common reasons — such as falls, chest pain, ankle sprain, or back pain — and hadn't used opioids in the six months before their hospital stay.

The study authors then split the group into patients who happened to be assigned to doctors who prescribed a lot of opioids, and patients who met doctors who didn't. It was a clever study design, since the major difference between the groups was now the doctors prescribed opioids, which allowed the researchers to figure out whether the prescribing patterns influenced patients' long-term use of painkillers. (In the study, "long-term use" was described as six or more months of daily opioids in the 12-month period following the ER visit.)

painkillers. (In the study, "long-term use" was described as six or more months of daily opioids in the 12-month period following the ER visit.)

painkillers. (In the study, "long-term use" was described as six or more months of daily opioids in the 12-month period following the ER visit.)

painkillers. (In the study, "long-term use" was described as six or more months of daily opioids in the 12-month period following the ER visit.)

painkillers. (In the study, "long-term use" was described as six or more months of daily opioids in the 12-month period following the ER visit.)

painkillers. (In the study, "long-term use" was described as six or more months of daily opioids in the 12-month period following the ER visit.)

59

x10,000

Article contents are highlighted by content categories you choose

... and the labels train AI via supervised machine learning

Archives



Digitized:

- Transcripts
- Meeting Minutes
- Letters/Correspondence
- News Reporting
- Magazines
- Literature
- Court Records

Labeling/Indexing



- Named Entities
- Event IDs
- Network Ties
- Faction Tracking
- Utterances
- Convo Analysis
- Opinion Analysis
- Topic Tagging & Clustering

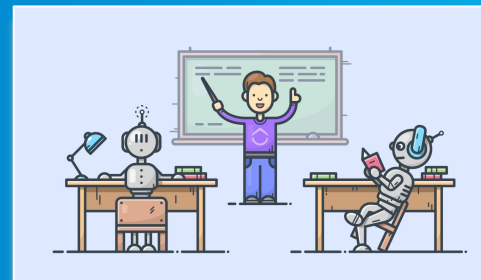
SCALE UP



Your Volunteers
MTurk Crowd

SciStarter
Volunteer Science
?Zooniverse?

Machine Learning



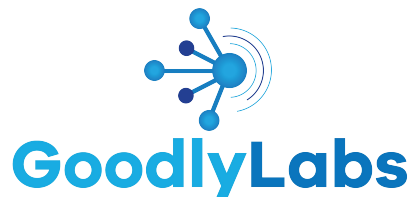
- Supervised ML
- Active Learning
- Human in the Loop Data-Validation
- Train AI to Label Archives Like Human Experts





Public Editor

A Project of:



Empowering individuals
with collaborative tools to
find common ground and
build a better society

A central hub of research and
education at UC Berkeley
designed to facilitate and nurture
data-intensive science



Powered by:



High-scale, expert-grade data
labeling for natural language

- End-to-end project management GUI
 - No command line, extra PMs, or data scientists required
- Data validation features for
- Research-grade data – backed by SAGE Publishing, global leader in social science methods

nick@goodlylabs.org – publiceditor.io

nick@thusly.co – tag.works

The future of work

“To make use of the strengths and limitations of ML, organizations will need to redesign workflows and rethink the division of tasks between workers and machines ... The resulting changes in work design will alter the nature of many jobs, in some cases profoundly. But the implications for specific skill groups are as yet uncertain and will in part depend on managerial and organizational choices, not on technologies alone.”

“The Work of the Future: Shaping Technology and Institutions”

Fall 2019 Report from the MIT Work of the Future Task Force

AI + ethics

“Technology design is the new policy-maker.” - Dr. Latanya Sweeney

“What came up time and time again [in AI training datasets] is the overrepresentation of lighter-skinned individuals, the overrepresentation of men, and the underrepresentation of women, and especially women of color ... Machines are learning from what? Data. So in this case, data is destiny.”
- Joy Buolamwini

“Race, Technology, and Algorithmic Bias,” Vision and Justice summit, Radcliffe Institute

Where can libraries lead?

1. What is AI good at, right now? Where does it struggle?
2. What is the role of data in AI/ML, and how can we procure, structure, document, and interpret data ethically for AI/ML use cases?
3. What does the AI-enabled organization look like, in terms of skill sets, workforce, business processes, and services?
4. How do libraries, as data stewards, work to debias datasets and promote an understanding of ethical application of AI among practitioners?
5. How do we make good decisions about AI/ML tooling in our own tech environments, and how will we determine, strategically, what (and how) we build / select / use?

Thank you!

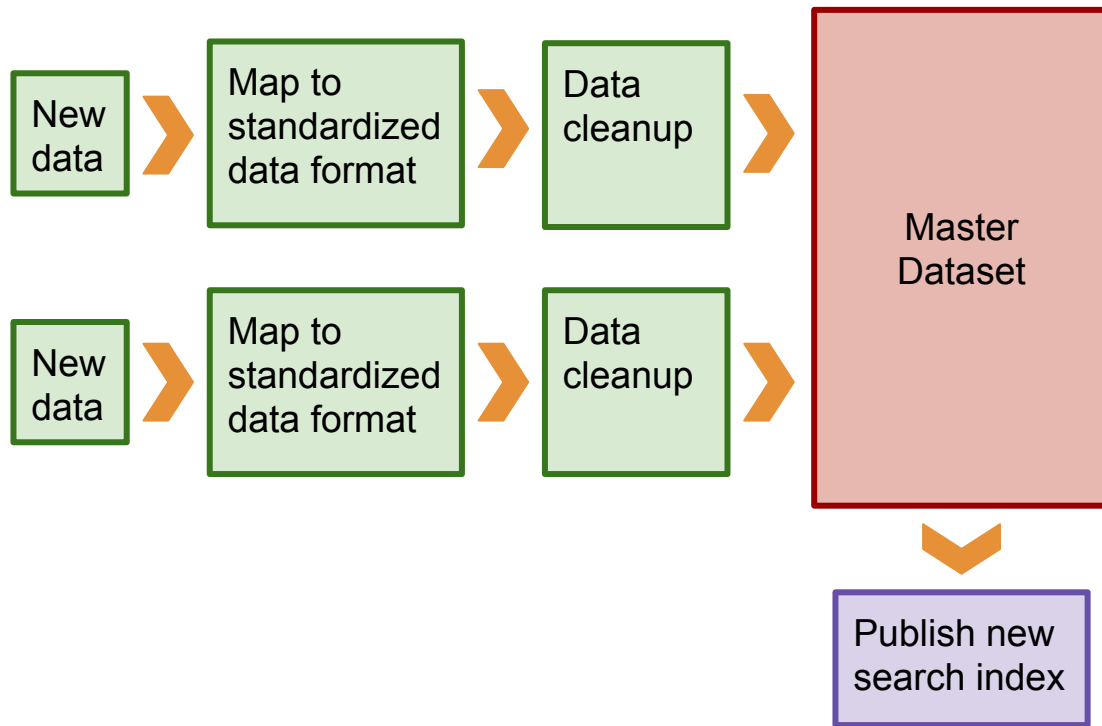
Heather Yager, Associate Director for Technology, MIT Libraries

Machine Learning at DPLA

1. The background

DPLA's new ingestion system

- Streamlined batch-processing workflow
- Cloud & cluster computing
- Standardized data model
- 35 million records (and growing!)



2. The challenge

Integrate ML into our production workflow

- Handle all of our data
- Handle regular data updates
- Ability to improve ML model over time
- Push-button simplicity
- Fast turn-around
- Use open-source tools
- Produce useful, usable output

3. The project

Recommendation system

Presbyterian Church in Pauls Valley, Oklahoma



[View Full Item](#)

Description

Photograph of the Presbyterian Church in Pauls Valley, Oklahoma.
Published by C.P. Bruce.

RELATED ITEMS



Maywood Presbyterian Church



First Presbyterian Church



First Christian Church



Wesley Methodist Church



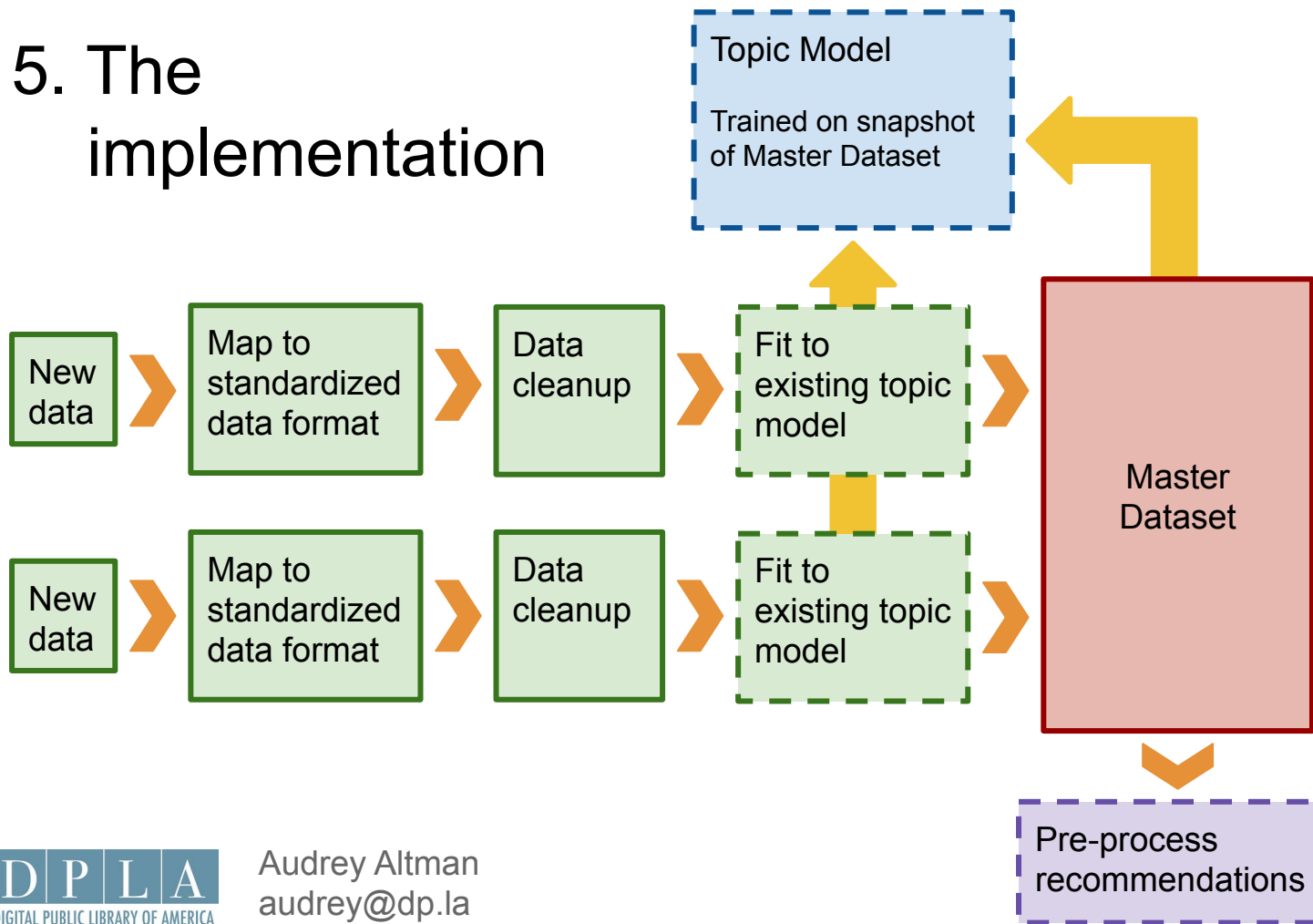
First Baptist Ch

4. The aspiration

If we can do this...

- what else can we do?
- how can we help other libraries do it too?

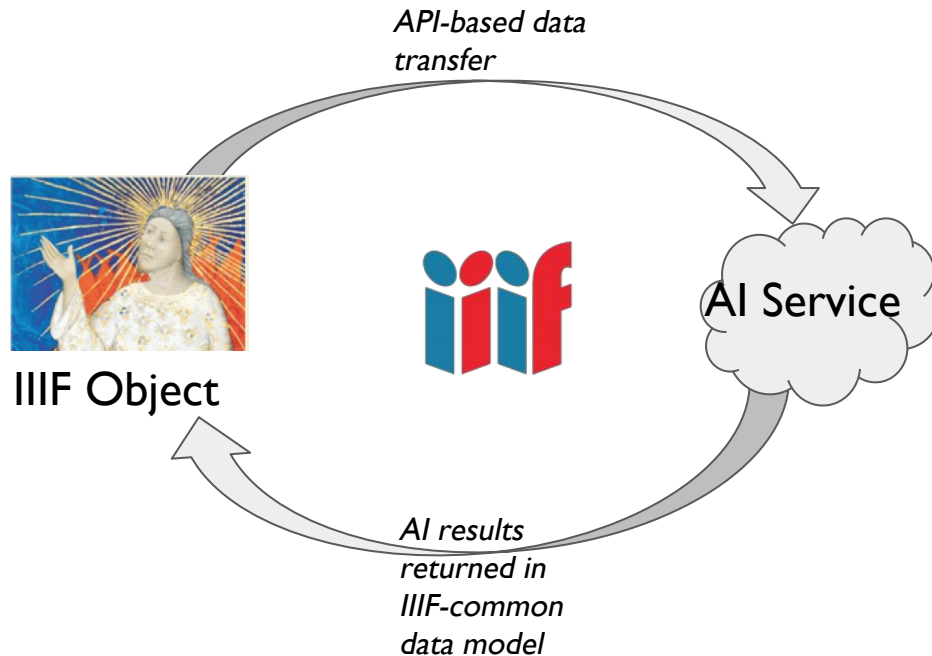
5. The implementation



IIIF and AI/ML: Chocolate and Peanut Butter



1. Easy access to corpora via IIIF
2. Access to any online AI service
3. Express results back as IIIF annotation lists
4. Assemble very large & cross-institutional corpora with IIIF



Examples:

- Bibliothèque nationale de France ([presentation](#); [example](#))
- National Library of Norway ([presentation](#); [example](#))
- National Library of Poland ([presentation](#))
- ROIS-DS Center for Open Data in the Humanities, Japan ([presentation](#); [example](#))

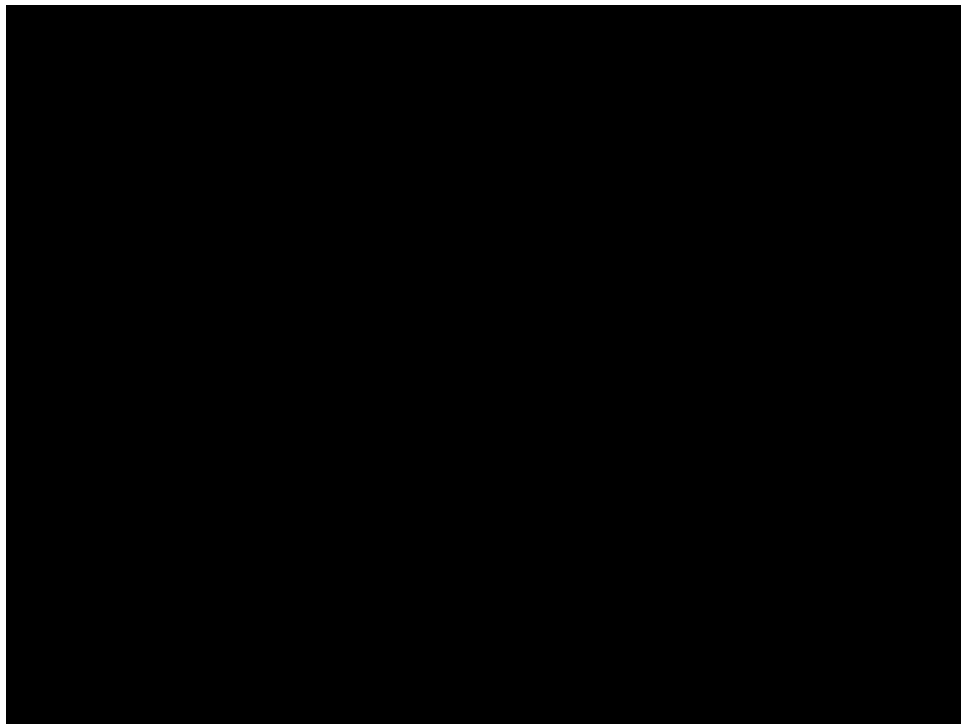
Example: AI OCR detection on Cursive Kuzushiji characters



Embedded within a IIIF viewer ([IIIF Curation Viewer](#))

Kuzushiji, a cursive writing style, appears in 3M+ books, and was for over a thousand years but the standardization of Japanese textbooks in 1900 means “most Japanese natives today cannot read books written or printed just 120 years ago.” ([CODH source](#))

Today, the CODH IIIF platform applies in-viewer AI OCR ([Example](#))



AI Efforts at Stanford | LIBRARIES

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF ELECTRICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Enhanced Cataloging for Theses & Dissertations

Question: Can we use AI to *automatically* add keywords & subjects to catalog records?

Model & Engine: Yewno

Key Points: Integration with traditional cataloging workflow & ILS



Computer Vision for Archaeological Photos

Question: Can we usefully classify & mine under-organized photos from a 20 year dig?

Model & Engine: Claif.ai, Google CV, AutoML

Key Points: Comparison of commercial vs. custom models, integration with dig records

<http://catalhoyuk.com/research/imageai>



Tom Cramer
AUL & Director,
Digital Library
Stanford Libraries
tcramer@stanford.edu



To identify and establish channels for concrete exchange among LAMs for practical developments and application of AI; this includes practical outcomes to build a common ground for AI in libraries, with emphasis on:

- 1. identifying & establishing how-to's and best practices*
- 2. sharing use-cases demonstrating the power and potential of AI in LAMs*
- 3. facilitating capacity development in institutions for AI investigation,*
- 4. experimentation and production services*
- 5. coordinating collaborative efforts across institutions and nations*
- 6. organizing contacts with relevant R&D & commercial organizations*
- 7. training community members through bootcamps, webinars, et al.*
- 8. understanding and advancing the application of AI to IIF*
- 9. to build & exchange data sets & models of interest among libraries*

- Fantastic Futures (AI Conference + Workshops), Dec 4-6, 2019
- Google Group: ai4lam@googlegroups.com (207)
- Slack: <https://bit.ly/ai4lam-slack> (116)



Tom Cramer
AUL & Director,
Digital Library
Stanford Libraries
tcramer@stanford.edu

Machine classification of volumes in the HathiTrust Digital Library



HATHI
TRUST

- The HathiTrust Research Center develops tools and services to facilitate computational text and data mining of the HathiTrust corpus, including:

Web-based tools | Datasets | Secure computing environments |
Collaborative projects with scholars

- Researchers have engaged in projects that use machine classification to derive insight about the HathiTrust Digital Library, English-language literature, U.S. publishing, and more.

Library-scale classification examples

“Stable Random Projection: Lightweight, General-Purpose Dimensionality Reduction for Digitized Libraries” - Benjamin Schmidt

- Used stable random projection (dimensionality reduction) of HathiTrust to classify volumes, create cluster visualizations of subject and genre, and demonstrate corpus alignment.
- [Read more](#)

“Page-Level Genre Metadata for English-Language Volumes in HathiTrust, 1700-1922” - Ted Underwood

- Employed classification algorithms to identify the broad genre (fiction, poetry, drama, nonfiction prose, paratext) at the page level for the HathiTrust’s public domain collection.
- [Read more](#)

Corpus-scale classification examples

“The Transformation of Gender in English-Language Fiction” - Ted Underwood, David Bamman, and Sabrina Lee

- Detected the adjectives most commonly associated with female characters in their corpus, and then used a classifier to determine how stable the language was over time.
- [Read more](#)

“How Capitalism Changed American Literature” - Dan Sinykin

- Compared novels published by large publishing companies with those published by non-profits to see if a machine could learn to distinguish between the two.
- [Read more](#)

Civil War Photo Sleuth:

Combining Crowdsourcing and Face Recognition to Identify Historical Portraits

Kurt Luther, Vikram Mohanty, Paul Quigley, Ron Coddington



Library of Congress

Freeman Mason, 17th VT
Infantry and Michael Mason
(inset), 6th VT Infantry
Library of Congress



“...card-portraits...as everybody knows, have become the social currency, the sentimental ‘green-backs’ of civilization...” —Oliver Wendell Holmes, July 1863

Estimated 4M Union portraits
survive today (Coddington)

Only 10-20% identified
(Vaughan/Zeller)

Historians, genealogists,
archivists, collectors, dealers
seek to identify unknown
portraits

- Correct the historical record
- Create economic value
- Recognize contributions of marginalized groups
- Make personal connections

Digital Archive of Reference Photos

PUBLIC COLLECTIONS

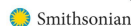
LIBRARY
LIBRARY
OF CONGRESS

THE
AMERICAN CIVIL WAR
MUSEUM
Confederacy ★ Union ★ Freedom



U.S. Army
Heritage & Education Center

National
Portrait
Gallery



NATIONAL
ARCHIVES

PRIVATE COLLECTIONS



How It Works

Visual Tags, Search Filters, and Face Recognition



Yale University Library

Microsoft
Azure



	Solon A Carter Ranks Held LTC MAJ CAPT Units Served US 14th NH Infantry Co. G
	Richard P DeHart Ranks Held LTC COL BRIG GEN Units Served US 46th IN Infantry US 99th IN Infantry US 128th IN Infantry
	James T Conklin Ranks Held LTC MAJ CAPT COL LTJ BRIG GEN Units Served US 14th WI Infantry

Outreach and Community-Building

SOCIAL MEDIA

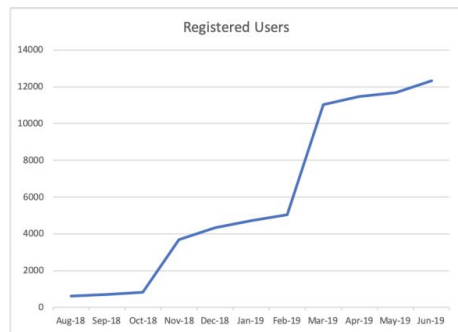


IN-PERSON EVENTS

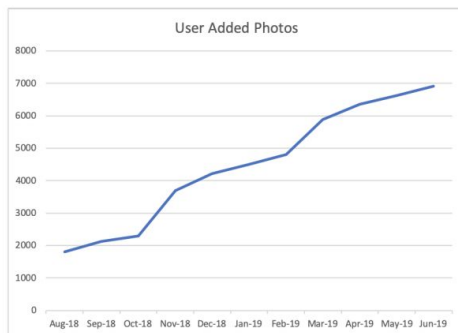


Growth Since August 2018 Launch

10,000+ REGISTERED USERS



7,000+ USER-CONTRIBUTED PHOTOS



28,000+ PHOTOS IN ARCHIVE

Initial Results and Next Steps

Identifications and Success Stories

PUBLIC COLLECTIONS



Francis M. Eveleth



New York
Public
Library

PRIVATE COLLECTIONS



William H.
Baldwin

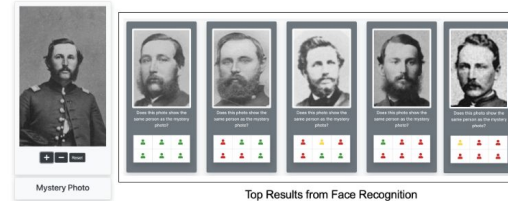


Future Work

BENCHMARKING



LAST-MILE PROBLEM



MIXED-INITIATIVE VISUAL TAGGING



Ross Goodwin

Not a poet.

1 the Road →

1theRoad.com

rossgoodwin.com



@rossgoodwin



@ross.good.win

ross.goodwin@gmail.com





"Please Feed The Lions"
// Es Devlin x Ross Goodwin

Trafalgar Square //
London Design Festival
// September 2018

<https://g.co/pleasefeedthelions>

A machine learning crowdsourced poem

by Es Devlin



narratedreality.com

Ichneumonis picturam hanc desumpsimus ex uera eius effigie cum Crocodilo nobis conspecta, Bellonius. Ex eodem postea cognoui dorsum in hac pictura nimis eleuari, & planius esse debere: rostrum paulò magis acuminatum, crura minùs crassa. Colorem, talem ferè esse, qualis hîc nigro alboq; distinctus conspicitur.

Networked Texts

Improved Inference by Exploiting Relational Structure

David Smith

NULab for Texts, Maps, and Networks
Khoury College of Computer Sciences
Northeastern University, Boston, MA

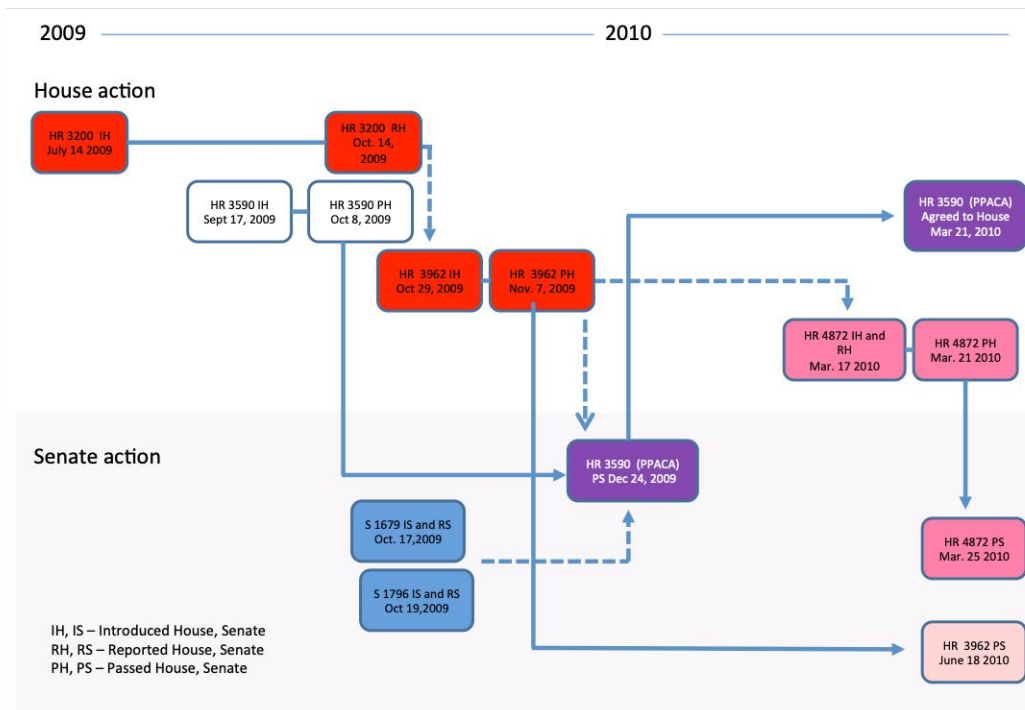


Northeastern University

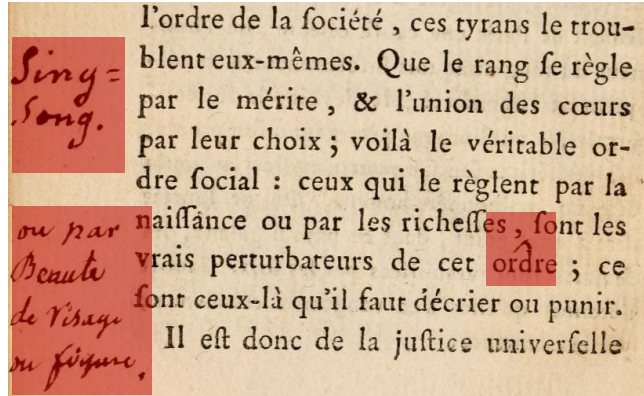
NULab
for texts, maps, & networks

Networked Texts

- Documents are not independent
 - We observe, e.g., many versions of a bill on its way to passage, as here with the Affordable Care Act
- We work on:
 - inferring links among documents;
 - exploiting links for better transcription, classification, information extraction, etc.

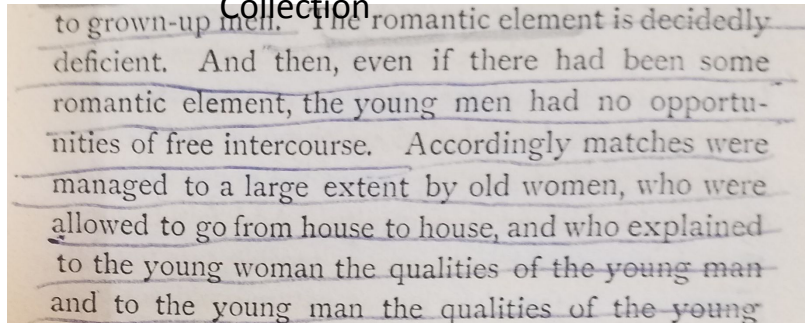


Variant Editions, Noisy OCR, Reader Annotations

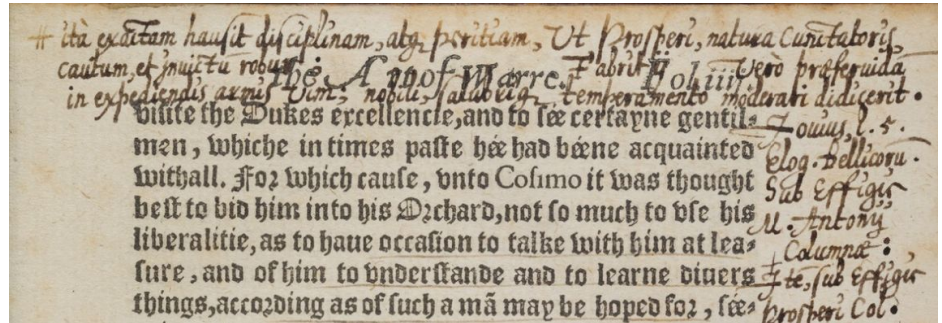


Tordre de la fociété , ces tyrans le trou-
Si^ct^ ^^ei^f ^"x-mêmes. Que le rang fe règle
r^_^ par le mérite , & l'union des cœurs
J par leur choix ; voilà le véritable or-
dre focial : ceux qui le règlent par la '
iTM/ /7a.r naiflTânce ou par les richesses , font \o
tuutu ^^^^^ perturbateurs de cet or3re j ce
J Y^ ^onr ceux-là qu'il faut décrier ou punir.
/, II eft donc de la iuftice univerfeile

BPL Adams
Collection



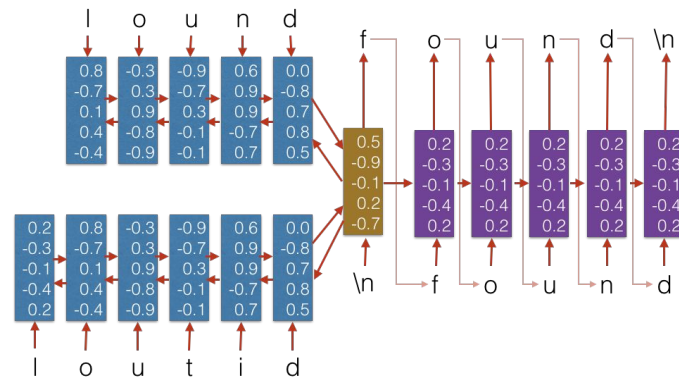
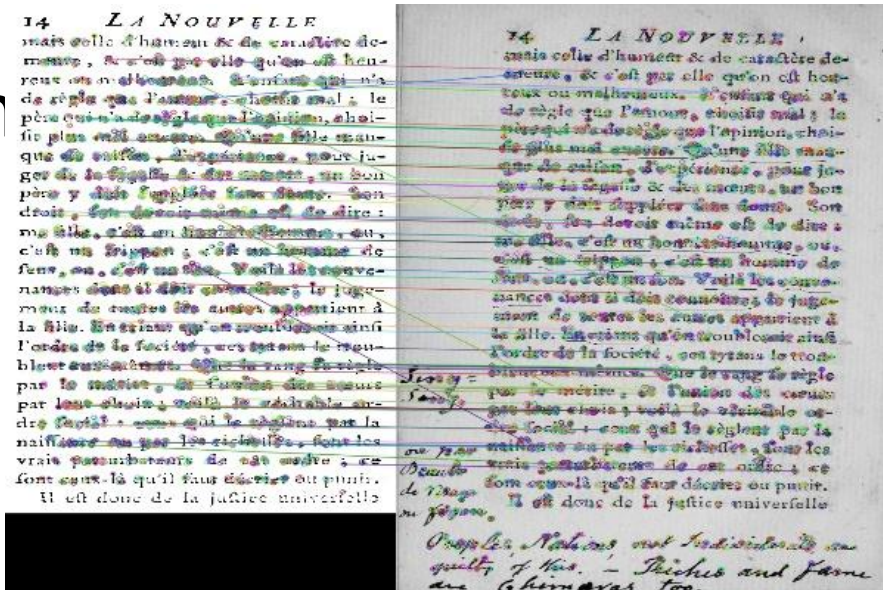
Book



Archaeology of

Ichneumon Approach

- Collate texts and images
- Infer consensus transcriptions w/multi-input attention
- Train OCR correction models to improve single outputs
- Train object localization models for annotation detection from single images
- **Data:** Cleaned OCR, database of annotation locations

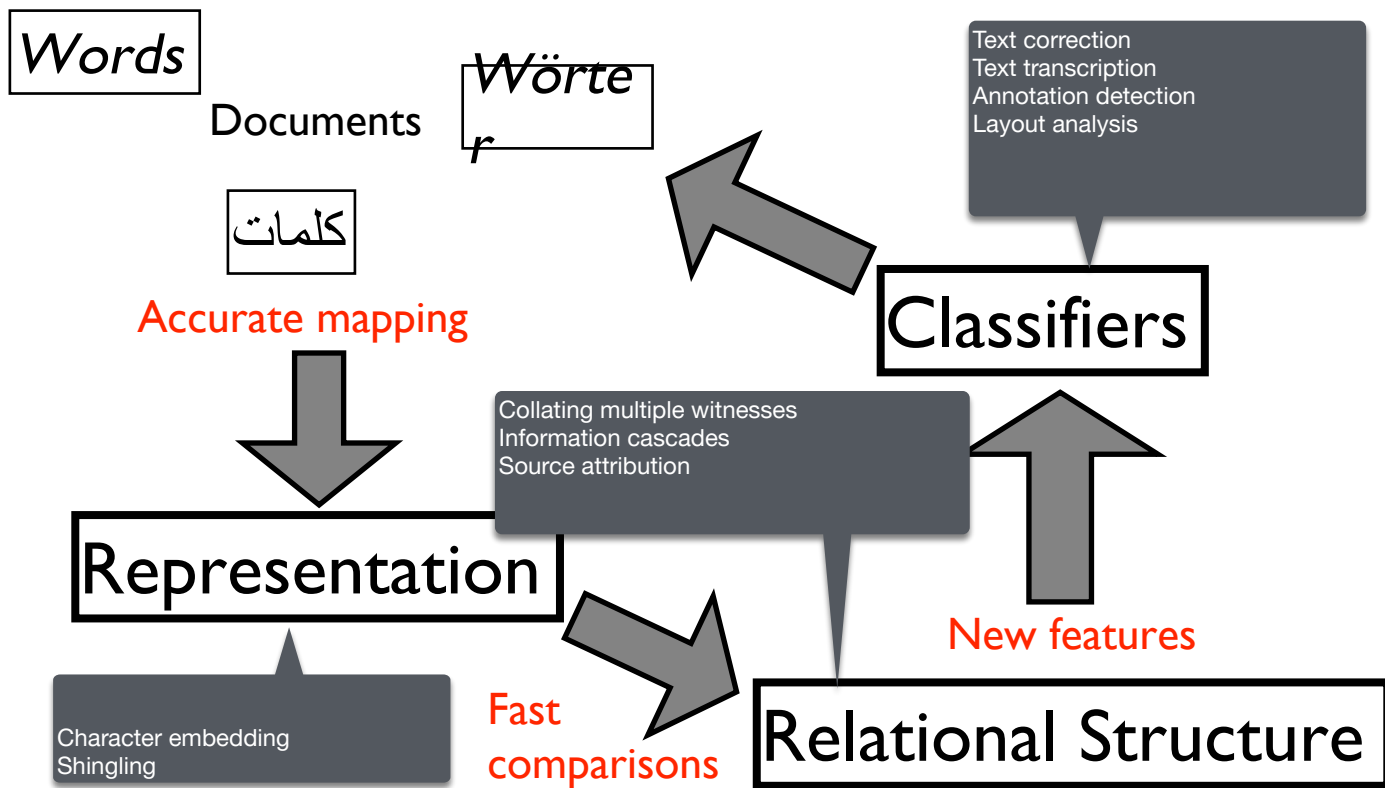


Arabic-script OCR

- Interactive training and post-correction of OCR models
- Computer-vision-based layout analysis models, trained on semantic markup in documents
- Training by exploiting existing scholarly and community-built editions
- Domain adaptation and model combination



Synoptic View of ML Applications



Related Projects at Northeastern's NULab



- **Viral Texts:** viral culture in the US
 - <https://viraltxts.org>
- **Oceanic Exchanges:** global information propagation
 - <https://oceanicexchanges.org>
- **Ichneumon:** collating books and detecting annotations
 - <http://www.ccs.neu.edu/home/dasmith/ichneumon-proposal.pdf>
- **KITAB:** knowledge, information technology, and the Arabic book
 - <http://kitab-project.org/>
- **Arabic-script OCR Catalyst Project**
 - <https://medium.com/@openiti/openiti-aocp-9802865a6586>



: A computational helper to describe digital images

ALL U
NEED



Harish Maringanti

Associate Dean of IT & Digital Library

Vivek Srikumar

Assistant Professor, Computer Science

Dhanushka Samarakoon

Assistant Head of Software Development

Bohan Zhu

Software Developer

Goals

Enhance discovery experience for users

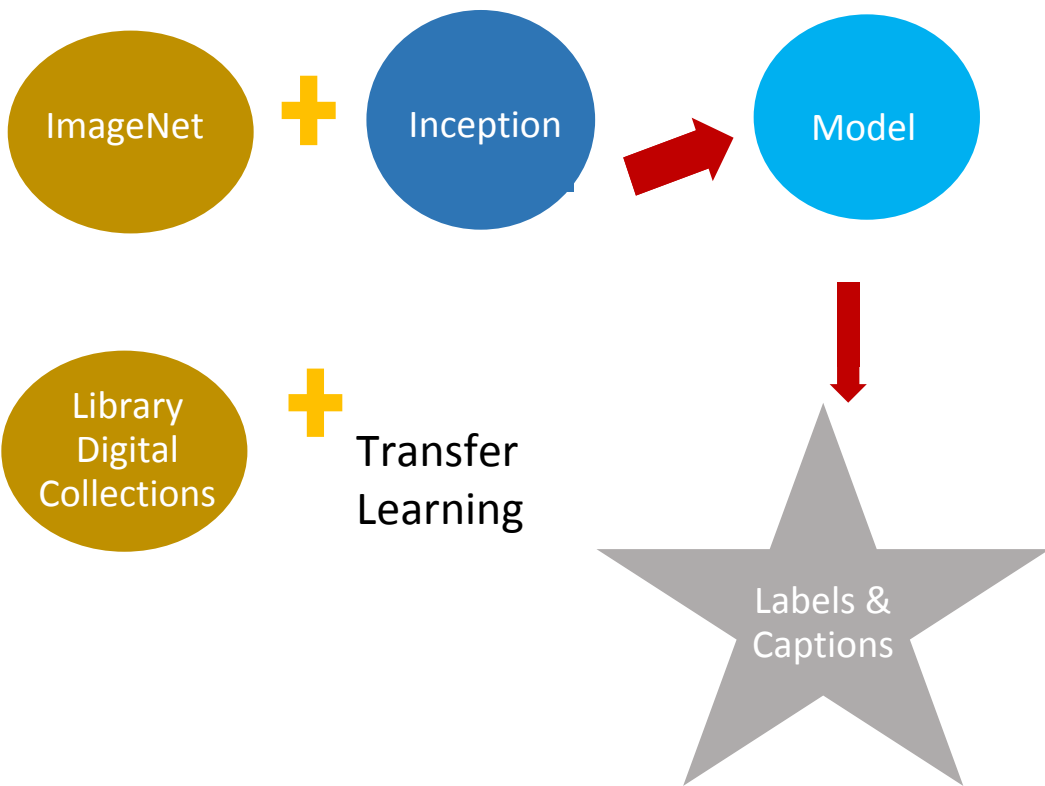
Expedite metadata creation

Address backlog issues in processing collections

Work on this project funded in part by



Preliminary Work



Issues

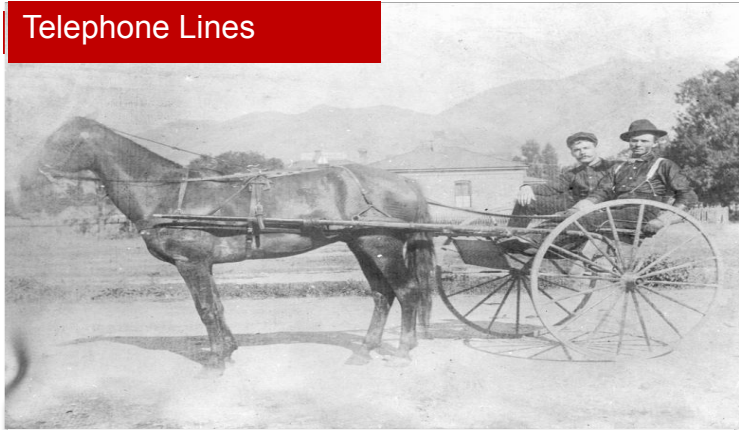
Domain Adaptation: A model that is trained to optimize predictive accuracy on one domain (e.g., general web images captured in cell phone cameras) may not be well-suited for another domain (e.g., archival scans of black and white photographs).



- a man and a woman sitting at a table. ($p=0.000579$)
- a man and a woman sitting in front of a laptop computer. ($p=0.000110$)
- a man and a woman sitting at a table with a laptop. ($p=0.000096$)

Issues

Telephone Lines



Black and white photo of horse drawn carriage

Student group portrait



Vestment; Hoopskirt; Crinoline; Suit; Clothing

Next Steps

Prototype of a tool that assists humans in metadata generation

Labels expand


sandbar sand bar speedboat

paddle boat paddle canoe seashore coast seacoast sea-coast

[CAN'T FIND WHAT YOU NEED?](#)

Tag: +

Caption



🔒 🗑️ 📊

a man riding a wave on top of a surfboard .

a person riding a surf board on a wave

a man on a surfboard riding a wave .